# Data-driven modeling of the neural dynamics underlying language processing

Julia Berezutskaya

ISBN: 978-90-393-7270-8

Cover: Max Philippi (maxphilippi.com)

Print: Ridderprint (ridderprint.nl)1

Copyright © 2020 by Julia Berezutskaya

# Data-driven modeling of the neural dynamics underlying language processing

## Datagestuurde modellering van de neurale dynamiek die ten grondslag ligt aan taalverwerking

(met een samenvatting in het Nederlands)

**Proefschrift**

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 14 april 2020 des ochtends te 10.30 uur

door

**Julia Berezutskaya**

geboren op 12 oktober 1990
te Berezniki, Rusland

## Promotoren:

Prof. N.F. Ramsey

Prof. M.A.J. van Gerven

## Copromotor:

Dr. Z.V. Freudenburg

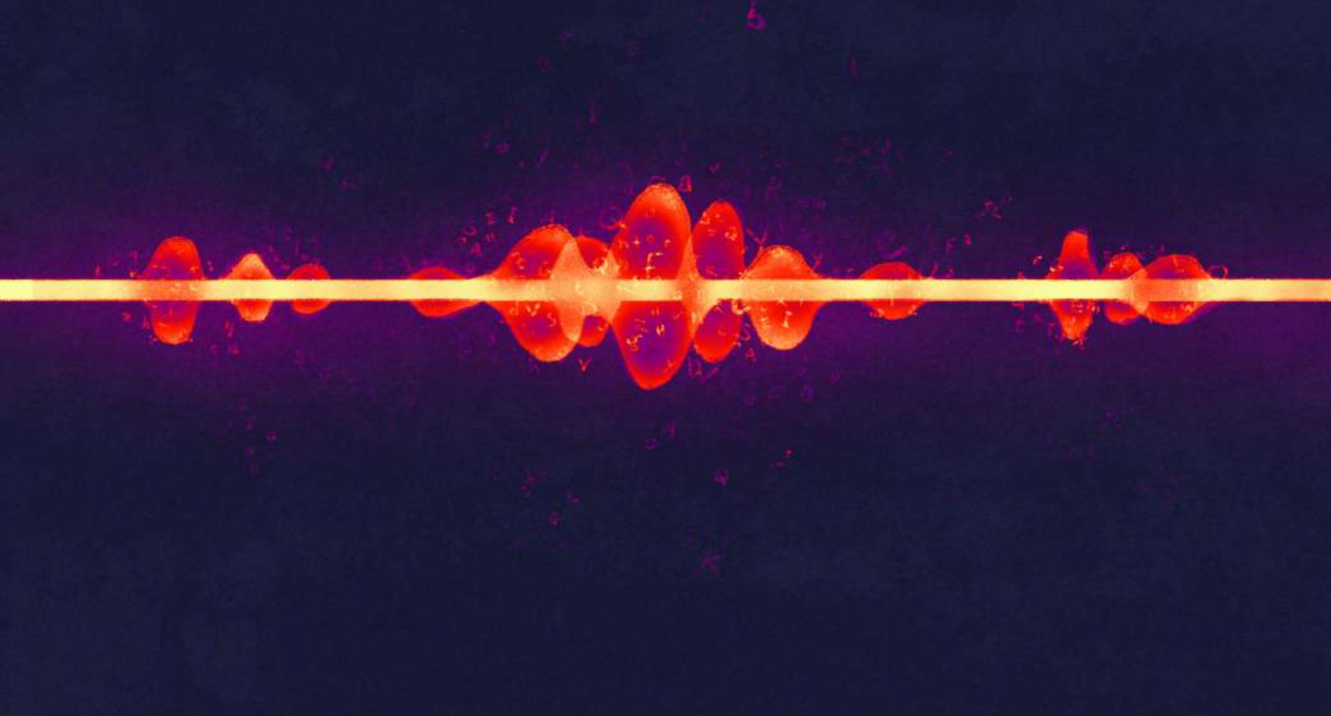
Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van Language in Interaction Project Gravitation Grant

## CONTENTS

chapter

# 1

# Introduction



# CHAPTER 1. INTRODUCTION

## 1.1 INTRODUCTION

Understanding the neural mechanisms underlying language perception is one of the key topics in cognitive neuroscience. Over decades of work, a lot of knowledge has been accumulated based on evidence from psycholinguistics, neuropsychology, neuroimaging, electrophysiology and many related disciplines. We have come to know about the core network of cortical regions involved in speech perception, fundamental components in the neural signal triggered by speech properties and the basic principles of the hierarchical organization in language reflected in the brain responses.

A lot of the work that has contributed to our understanding of speech processing in the human brain comes from hypothesis-driven experiments that rely on controlled experimental paradigms. Although this research undoubtedly enriched our understanding of the topic, alternative, more data-driven approaches have recently begun to interest neuroscientists. On the one hand, data-driven methods can allow us to go beyond experiments with controlled, limited language material and focus on perception of continuous naturalistic speech as when listening to a story, watching a film or during a real-world conversation. On the other hand, using data-driven techniques may offer a new way to study the neural responses: instead of testing a concrete theory-based hypothesis on the neural data, these methods may allow for exploration of the neural responses and generation of ideas about the observed differences and patterns. Thus, the theory-driven approaches may focus more on assuming what kind of information should be captured in the brain patterns and actively look for its encoding, whereas the data-driven approaches may focus more on exploring the neural data itself and extracting optimal neural triggers in the speech and language stimuli in a bottom-up way.

Recent advances in computer technology and the rise of big data analytics have led to an increased interest in application of machine learning techniques to many fields of science, including cognitive neuroscience. Next to the type of analyses that have long been considered a standard practice in the field (such as generalized linear regressions, for example), machine learning offers a plethora of techniques to investigate the brain responses in a more unconstrained and bottom-up manner. For example, multiple unsupervised learning approaches can be employed to extract structures and patterns inherent in the brain responses to speech and language stimuli (clustering and data compression algorithms). Another class of powerful approaches can be used to extract data-driven speech and language properties that could trigger the neural responses (task-optimized artificial neural networks). Finally, variants of these techniques can be developed such that the speech and language properties they extract are optimized by the neural data itself (artificial neural networks trained to fit the brain data directly). These techniques are particularly suitable for analyses that involve unconstrained continuous speech that has not been segmented into theory-defined language labels.

In this thesis, we use data-driven techniques in an attempt to explain the neural data during speech and language perception in an unconstrained naturalistic setup. We aim to use a combination of data exploration approaches and data-driven feature extraction models to provide a more bottom-up way of studying the neural responses. We hope that these data-driven strategies can confirm some of the previous theory-based neurolinguistic findings but at the same time offer new insight regarding how the human brain processes speech related information.

### Hierarchical structure of language and speech

Perception of speech is a difficult topic because speech itself has a complex multi-level organization. When we listen to other people talk, we focus on the meaning of what they say. However, extraction of this meaning, or *semantic* information, from the perceived auditory input is a complex process. Semantics (the meaning of spoken language) is composed of meaning of individual words, details of the structure of the sentences, intonation contour, phrasal patterns and various extralinguistic information, such as gestures, facial expressions and other context.

As expression of language, speech contains important linguistic elements. Apart from semantics, it has a rich *syntactic* component that represents connections between individual words in a sentence. Semantic information is based largely on the *lexical* component, or meaning of individual words. Words are typically not considered the most basic building blocks of language. Words can be broken down into smaller meaningful parts – *morphemes* (such as word *roots*: ‘build-’ in ‘building’, ‘grow’ in ‘regrowth’ and *affixes*: ‘pre-’ in ‘pre-historic’, ‘-ful’ in ‘beauti-ful’ and ‘-ed’, in ‘play-ed’). Morphemes consist of the smallest units of linguistic information – *phonemes*: ‘p’, ‘a’, ‘h’, etc. On the perceptual level, the phonemes are typically represented by orthographic symbols or acoustic sounds. Altogether, these components form a hierarchical organization of language: top components delivering composite meaning are constructed from lower-level linguistic units. Thus, the complexity of the semantic information is built up based on the units down the linguistic hierarchy with units at the top expressing composite semantic information (sentences), and units at the lower levels expressing primitive (morphemes) or almost no semantic information (phonemes).

### Neurobiology of speech perception

The knowledge about individual language components and speech properties has guided extensive research on the neural mechanisms underlying speech perception. Much of what we currently know about the field comes from work on neuropsychology, electrophysiology and neuroimaging. First evidence about the relationship between neural activity and language comes from lesion work of Paul Broca and Carl Wernicke, who outlined distinct sites in the human brain that could accommodate speech production (Broca, 1861) and speech perception (Wernicke, 1874). Later research has shown that even though much of the classical view on the functionality of these regions can be debated (D'Ausilio et al., 2012; Flinker et al., 2015), it was a crucial step in beginning to uncover the neural underpinnings of language.

Further work focused on the neural responses to individual language units, such as phonemes, words or sentences; and responses during different aspects of linguistic processing: phonological, lexical, semantic and syntactic processing. Specifically, various studies compared the neural responses to individual phonemes, words and sentences to find spatial and temporal differences associated with each language unit. Neural responses to different syntactic processes and semantic distinctions were also typically compared. As a result, a network of core cortical regions involved in speech perception was identified. These regions involve superior temporal, inferior frontal, middle temporal and precentral gyri (Friederici, 2012). Depending on the perceived modality (the nature of the input signal), primary auditory, visual or sensorimotor cortices are involved. Each of the core regions is typically associated with a range of functions, although many details of these regions' functionality are still widely debated. Thus, the posterior superior temporal cortex is known to support *phonological* processing (Binder et al., 2000; Indefrey and Levelt, 2004, Hickok and Poeppel, 2007), whereas the anterior superior temporal cortex is often associated with *supraphonemic* and word-level processing, as well as syntactic processing. The inferior frontal gyrus is involved in both semantic and syntactic processing (Friederici, 2011). The middle temporal gyrus is implicated in word retrieval and lexical processing (Dronkers et al., 2004; Krieger-Redwood and Jefferies, 2014). Involvement of the precentral gyrus (site for motor-related activity) is typically explained through engagement of speech motor cortices in articulatory mapping and prediction during speech perception (Galantucci et al., 2006; Liberman et al., 1967).

Several neurolinguistic theories have been developed based on the mapping of language components onto the brain responses during speech perception. These theories are based on the anatomical distribution of the neural responses to various language tasks, the temporal components in the neural responses (such as N400 and P600) and the structural and functional connections between the anatomical regions. One influential theory is the *dual-stream theory* inspired by the success of the similar framework in vision (Ungerleider, 1982). The dual-stream theory of speech processing (Hickok and Poeppel, 2004, 2007) shows a mapping of different language components onto the brain areas through two neural pathways: the ventral stream that subserves lexical and conceptual processing and the dorsal stream that subserves auditory-to-motor mapping. Another theory that places a large focus on the motor aspect of language in speech perception is the *motor theory of speech perception* (Liberman et al., 1967; Liberman and Mattingly, 1985). The theory suggests that similar neural mechanisms underlie both speech perception and speech production. The m*emory, unification and control theory* (Hagoort, 2005, 2013) emphasizes that various non-language specific components are engaged and interact during language perception. It states that speech perception is inevitably embedded in other cognitive systems and urges to study language in interaction with them.

The principle that unifies these theories along with the majority of works that have been done so far is based on focusing on known language constructs and processes, and searching for their representation in the neural responses. This is often achieved by comparing the brain activity in different language conditions (for example responses to phonemes and responses to individual words) relying vastly on the principle of *cognitive subtraction* (Donders, 1969). Thus, this research is largely influenced by (1) our theoretical views on the language system and the concepts we use to describe it and (2) the controlled experiments with limited language material aimed at isolating language units and processes of interest.

## Data-driven approaches to study the neural responses during speech perception

Undoubtedly, we have learned a lot from the traditional methods and theories that have been developed to relate individual aspects of speech perception and language organization to brain responses. At the same time, a lot remains debated or unknown about the neural organization of speech processing. Therefore, we think that the field could use an approach that is alternative to the theory-driven neurolinguistic research. Instead of assuming and searching for predefined language constructs and processes in the neural data, the field could benefit from methods applied to the neural responses to unconstrained continuous language material to explore the inherent patterns in the neural data and observe how the language related structures and organization arise from the brain responses directly. Given the recent rise of machine learning, big datasets and data-driven approaches, we believe that we now have the opportunity to apply these concepts to studying the neural mechanisms of continuous speech and language perception. Thus, instead of assuming a relationship between well-defined language units, such as phonemes, words and sentences, and trying to find their representation in the neural responses, we can focus on exploration of the brain patterns, extracting structures from them and interpreting these structures in the context of speech processing post-hoc.

A number of studies have already attempted this approach and used data-driven techniques to enrich our understanding of the neural processes underlying cognition. These works show that data-driven techniques provide results that not only agree with existing theoretical views but offer new detail regarding how the brain processes information. For example, there is now data-driven evidence that not only is the hierarchical organization inherent in human behavior and external stimuli, but it is also reflected in the structure of the neural responses during perception (Güçlü and van Gerven, 2015a; Kell et al., 2018), memory processing (Hasson et al., 2015) and action planning (Grafton and de C. Hamilton, 2007). In visual perception, using data-driven visual features the researchers showed a gradient of increasingly complex semantic information propagating from early visual cortex to high-level associative temporal and fusiform cortices (Güçlü and van Gerven, 2015a). In processing of language, there has been data-driven evidence that human brain uses linguistic context and word co-occurrence patterns to encode meaning of individual words (Mitchell et al., 2008).

In this thesis, we extend the use of these data-driven approaches to studying the neural representation of language and speech in an unconstrained naturalistic setup. This work follows a less traditional route and combines three important components that should maximize the success of our approach. First, the projects make heavy use of data-driven techniques aimed at data exploration as well as feature extraction for fitting neural encoding models. Second, the neural recording modality is electrocorticography, which we believe has the necessary quality as well as high temporal and spatial resolution for achieving best results. Third, we opt for naturalistic experimental paradigms, which provide the least number of constraints on language and speech perception. Thus, subjects are meant to process speech in a way similar to non-experimental, every-day situations without being restricted to only some aspects of language processing that are of interest to the researcher.

## Overview of data-driven approaches used in the thesis

Several machine learning approaches are used extensively throughout the thesis. We make a distinction between data exploration and data fitting models. In both cases we assume that we have a set of neural recordings $R$ collected as a response to an external set of stimuli $S$. Then, we can choose (1) to explore the structure of $R$ and $S$ using data exploration methods and interpret their internal structures with respect to each other, and/or (2) to fit the neural responses $R$ to the properties of the stimuli $S$ using a combination of feature extraction and neural encoding models. Below we briefly explain the two types of approaches: (1) the data exploration approaches, for example, clustering and principal component analysis, and (2) feature extraction and neural encoding models, including multiple linear regression and artificial neural networks.

## Data exploration

### *Correlation and representation similarity analysis*

Among the simplest approaches aimed at exploring the data is the investigation of its correlation structure. Various correlation coefficients can be computed based on the underlying assumptions about the nature of the data. Here, we extensively use either Pearson (1.1) or Spearman correlation coefficient (1.2):

$$r_p = \frac{\mathrm{cov}(\mathbf{y}_a, \mathbf{y}_b)}{\sigma_{\mathbf{y}_a}\sigma_{\mathbf{y}_b}} \tag{1.1}$$

$$r_s = \frac{\mathrm{cov}(r\mathbf{y}_a,\ r\mathbf{y}_b)}{\sigma_{r\mathbf{y}_a}\sigma_{r\mathbf{y}_b}} \tag{1.2}$$

where $\mathbf{y}_a$ and $\mathbf{y}_a$are the vectors of neural activity for neural populations $a$ and $b$ respectively (for example, as measured by two electrocorticography electrodes); $\sigma_{\mathbf{y}_a}$ and $\sigma_{\mathbf{y}_b}$ are the standard deviations of the neural activity vectors; $r\mathbf{y}_a$ and $r\mathbf{y}_b$ are rank-transformations of the neural activity vectors, and $\sigma_{r\mathbf{y}_a}$ and $\sigma_{r\mathbf{y}_b}$ are the standard deviations of the rank-transformations. The neural activity can be measured as response to a certain task or stimuli, but can also represent neural recordings that were not modulated experimentally.

Representational similarity analysis (RSA, Kriegeskorte et al., 2008; Nili et al., 2014) makes extensive use of data correlation patterns by comparing the correlation structure across various data sources. In particular, the comparisons can be performed across participants, neural recording modalities, behavioral responses and stimuli representations. Conceptually, the method aims at unraveling patterns in the inner structure of a target dataset (for example, neural activity dataset) and assessing the similarity in the inner structure of this target dataset with multiple candidate ones (for example, various *feature representations* of the stimuli). The method has been used expensively in cognitive neuroscience research (Cichy et al., 2014; Groen et al., 2018; Khaligh-Razavi and Kriegeskorte, 2014).

Thus, each dataset (target neural dataset and candidate feature datasets) is represented as a matrix of size $n \times m_d$, where $n$ is a number of time points (for example, trials or stimuli) and $m_d$ is either the activity of $d_1$ neural populations responding to one trial/stimulus (target neural dataset) or the representation of the same trial/stimulus by $d_2$ features in a stimulus feature set (candidate dataset). Thus, in RSA $m_d$ can be different across datasets, which makes it a powerful method to compare similarity of representations varying in dimensionality. As a first step in RSA, representational dissimilarity matrices (RDMs) are calculated per dataset. This yields symmetrical matrices of size $n \times n$ containing pairwise dissimilarities (1 – correlation) in each dataset. Each value represents a dataset specific dissimilarity for each pair of trials/stimuli. As a second step, the upper triangular part of the target RDMs is correlated with the upper triangular part of each candidate RDM to assess the similarity of the inner structure of the neural responses (target dataset) with the various feature sets (candidate datasets). In order to assess the correlation between target and candidate representations statistically, multiple correlation instances for each target-candidate RDM pair need to be computed. This can be achieved by computing target-candidate RDM correlations over multiple subjects or runs of the experiment. Finally, for each target-candidate pair the significance of their overall correlation $\hat{\rho}$ is assessed using a non-parametric one-sided Wilcoxon signed-rank test:

$$w = \sum_{i=1}^{k} [\mathrm{sign}(\rho_i) \cdot r_i] \tag{1.3}$$

where $k$ is the number of correlation instances (for example, subjects or runs), $\rho_i$ is the target-candidate correlation per instance $i$ (for example, per subject or run) and $r_i$ is the rank of the correlation $\rho_i$. The value of $w$ statistic is compared against a critical value under a defined significance threshold. For large values of $k$, a z-score statistic $z$ is computed based on the normal approximation.

The difference in the degree of similarity to the brain responses (target dataset) between two or more feature sets (candidate datasets) can also be assessed statistically using a two-sided Wilcoxon signed-rank test comparing two or more sets of correlations, each for a specific target-candidate RDM pair.

### *Affinity propagation clustering*

Another approach to exploring the inner structure of data is clustering. There exist a variety of clustering approaches, each being sensitive to specific properties of the data. Various clustering techniques have been applied to the brain data, and here we focus on a clustering approach called affinity propagation (AP, Frey and Dueck, 2007). AP clustering has a number of benefits compared to other methods. First, AP clustering is non-parametric, which means that it does not require specification of a number of clusters prior to running it. Instead, the number of clusters is determined automatically based on a *preference* parameter, that estimates how likely each data point is to be a center of a cluster. Another benefit of AP clustering is identification of clustering *exemplars*, which means that the most representative data points become cluster centers. This is convenient for the interpretation of the clustering results as no additional transformation between the clustering outcome and the original data is required. Thus, for example in case of clustering the brain data (such as activity of electrocorticography electrodes), we can simply investigate the time course and the spatial localization of the cluster exemplars to get an idea about the cluster's possible function.

AP clustering procedure is an iterative process, during which for each pair of data points two values are calculated and updated on each iteration: responsibility $r_{ik}$ (estimation of how well the data point $x_k$ is suited to be a cluster exemplar for data point $x_i$) and availability $a_{ik}$ (estimation of how appropriate it will be for data point $x_i$ to pick data point $x_k$ as its exemplar):

$$r_{ik} \leftarrow s_{ik} - \max_{k \prime \neq k} (a_{ik'} + s_{ik'}) \tag{1.4}$$

$$a_{ik} \leftarrow \min(0, r_{kk} + \sum_{i' \notin \{i,k\}} \max(0, r_{i\prime k})) \text{ for } i \neq k \tag{1.5}$$

$$a_{kk} \leftarrow \sum_{i \prime \neq k} \max(0, r_{i\prime k}) \tag{1.6}$$

where $k'$are data points other than $k$, that are also candidates to be a cluster exemplar, $s_{ik}$ is the similarity between data points $x_i$ and $x_k$. Data point similarity (or *affinity*) can be calculated as a distance metric (such as negative Euclidian distance) or a correlation metric (such as Pearson correlation coefficient).

The preference parameter $\vartheta$ regulates the resulting number of clusters, such that higher values of $\vartheta$ will lead to more cluster exemplars, and lower values of $\vartheta$ will result in less clusters. For example, if Pearson correlation ($\rho$) is used to calculate the similarity matrix, the highest value of $\vartheta$ would be $\max \rho \Rightarrow 1$. Unless there are duplicate data points, this value will only appear along the diagonal of the similarity matrix forcing each data point to become a cluster exemplar. If $\vartheta$ is lowered to $\min \rho \Rightarrow -1$, then all data points will be likely collapsed in one cluster as there is more tolerance on similarity within a cluster and therefore less data points are likely to be exemplars. Typically, a median or a mean of the similarity matrix is recommended as a starting choice for the preference parameter. The optimal value of $\vartheta$ can also be determined in a data-driven way using an 'elbow' method (= bend of the curve associated with optimal number of clusters), silhouette metric or information criteria (Akaike or Bayesian information criteria).

Both responsibilities and availabilities are updated on each iteration of the clustering procedure until the values stabilize over a number of iterations or until a total number of iterations is reached. Then, the cluster exemplars are defined based on the sum of availabilities and responsibilities ( $r_{ik} + a_{ik}$) of each data point $x_i$ and all exemplar candidates $x_k$:

$$e_i = \begin{cases} x_i & \text{if } \underset{k}{\operatorname{argmax}}(r_{ik} + a_{ik}) = i \\ x_k & \text{otherwise} \end{cases} \tag{1.7}$$

where $e_i$ is the exemplar for data point $x_i$.

Per data point only one cluster exemplar is selected, and data points that share the same exemplar are clustered together.

The clustering can be applied directly to the neural data to find different patterns of activity and relate the groups to stimulus properties or anatomical locations of the clustered neural sources. Clustering approaches can also be applied to stimuli properties to find structure in these data, or to the matrix of weights that map stimulus properties to the neural responses (**B**-weights of the neural encoding model described below).

### *Principal component analysis*

Another set of techniques that are useful for exploring the data are based on data transformations. Principal component analysis (PCA) is a linear transformation of multidimensional data. It helps compress the data by organizing them along several new dimensions, or axes, that are orthogonal to each other and are sorted in the order of decreasing variance in the data. Because there is no linear dependence between the new individual dimensions, each axis captures an independent and often meaningful representation of the data. PCA projects the original data matrix **X** to a new orthogonal space **T** by computing the weights of the transformation **W**:

$$\mathbf{T} = \mathbf{XW} \tag{1.8}$$

The weights are calculated such that the variance of **T** is diagonal and the elements along the diagonal are sorted in the descending order. One way to perform PCA is to compute the eigenvalue decomposition on the covariance matrix cov(**X**). Then the eigenvectors of this decomposition sorted by the magnitude of eigenvalues will represent the weights **W** for the projection of **X** to a new orthogonal space **T**.

PCA has been used extensively in various fields including applications to neural data (Andersen et al., 1999; Friston et al., 1993; Halai et al., 2017). It is also often used as a preprocessing step aimed at dimensionality reduction (Partridge and Calvo, 1998) and denoising (Zhang et al., 2010).

## Feature extraction and neural encoding models

### *Linear neural encoding models*

An approach to mapping of the stimulus properties onto the observed neural responses that has gained recognition in the recent years is the neural encoding model (Kay et al., 2008; Naselaris et al., 2012). Typically, the neural encoding model is a linear mapping from the stimulus properties to the brain response, thus it takes a form of a multiple linear regression. The weights of the regression along with other parameters (for example, regularization) are computed either using a closed form solution or through optimization algorithms, such as a gradient descent aimed at minimizing the *cost function* (typically, mean least squares). The quality of the fit is then estimated on a held-out test set, whose data was not part of training. This estimation also provides information about how well the model generalizes to unseen data and can therefore indicate existence of a valid stimulus-to-brain mapping.

$$\mathbf{y}_a = \boldsymbol{\beta}_a{}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\varepsilon}_a \tag{1.9}$$

where $\mathbf{y}_a$ is a vector of response to stimuli recorded from a neuronal population $a$ (for example as measured with an electrocorticography electrode), $\mathbf{X}$ is a matrix of stimuli properties, or *features*, and $\boldsymbol{\varepsilon}_a \sim \mathcal{N}(0, \sigma^2)$ representing the noise component. The linear neural encoding model then provides an estimate of $\boldsymbol{\beta}_a{}^{\mathrm{T}}$ which contains the mapping from the stimulus feature space onto the neural responses.

The stimulus properties passed as input to the neural encoding model can vary in their degree of complexity and can capture specific aspects of the stimulus. Thus, many alternative feature sets could be used to fit a neural encoding model onto the same neural data. Then, each feature set could fit only the relevant part in the neural responses. For example, based on our knowledge about the hierarchy of visual neural processing (Van Essen et al., 1992), low-level visual features would only fit early visual cortex responses, whereas object categories would fit responses in inferior temporal cortex.

### *Artificial neural networks for feature extraction*

The results achieved with linear neural encoding models depend to a large extent on the features chosen to represent the stimuli. A specific class of methods has recently gained a lot of attention in connection with extraction of complex multi-level features directly from unprocessed input data. Artificial neural networks (ANNs) have achieved success in many areas of applied research due to their top performance and applicability to a wide range of problems including automatic object recognition (He et al., 2016; Simonyan and Zisserman, 2014), text generation and analysis (Devlin et al., 2018; Karpathy and Fei-Fei, 2015), classification of medical images and many others. The success of ANNs is largely due to the recent advances in graphical unit processing technology as well as access to increasingly large amounts of data. Importantly, ANNs have been claimed to achieve human-like performance and make understandable errors that resemble those made by humans (Stallkamp et al., 2012; Taigman et al., 2014). This fact led to an increased interest in the features that ANNs extract from the input data by suggesting that human brains might be extracting similar representations when processing similar data, for example images or texts. Thus, ANNs have drawn considerable attention of cognitive neuroscience.

From the technical point of view, ANN is a mapping function $\varphi$ that projects the input data **X** to the output labels **l**:

$$\mathbf{l} = \varphi(\mathbf{X}) \tag{1.10}$$

As a typical example, this could represent a function mapping image pixel data to the categories of objects present in the images. This function can be linear and thus subserve a linear mapping. In its basic form a linear ANN is equivalent to a multiple linear regression. Multi-layered linear ANNs can provide useful information about the distinct linear components of the input. However, the real power of the ANNs lies in their capacity for a non-linear mapping, granting them the ability to approximate any relationship between input and labels, and thus making them *universal approximators* (Hornik, 1991). Structurally, ANNs are organized in layers and are said to have an input layer, an output layer and *hidden layers*. Hidden layers refer to computational modules that process intermediate representations within the ANN. A *shallow* ANN has only one hidden layer, which means that the input is passed into one computational module, then its output is used directly to produce the prediction of the label. Stacking several hidden layers together makes a multi-layered, or *deep*, ANN. Mathematically, both shallow and deep ANNs are universal approximators, however in practice, deep neural networks are faster and more efficient (Lin et al., 2017; Mhaskar et al., 2016). Moreover, the deep, multi-layered structure of an ANN lends itself to incorporating hierarchical relationships between features at different hidden layers, which is often desired and arguably leads to extraction of more intuitive interpretable features.

The architecture of deep ANNs is one of its defining properties. ANN architectures vary depending on the problem it needs to solve and the type of data it needs to process. The ANN architecture types include feed-forward ANNs, recurrent ANNs, ANNs for unsupervised learning, generative and discriminative ANNs. In this thesis, we will use a limited number of computational mechanisms for construction of the hidden layers. Specifically, we will make use of linear, convolutional and recurrent connections, each followed by a non-linear *activation function*. Based on these types of connections we will discuss three basic ANN architectures: multi-layered perceptrons, convolutional neural networks and recurrent neural networks. Multi-layered perceptrons (MLPs) are feedforward non-linear neural networks. MLP hidden layers are the linear transformations of the input followed by a non-linear activation function (for example, rectified linear units). MLPs were the first ANNs organized in layers, and thanks to the non-linear activation functions they were able to solve various complex tasks. However, specific problems proved to benefit from other types of computational mechanisms. *Convolutional neural networks* (CNNs) were the pioneer ANNs in the field of automatic object recognition and were originally trained to associate images with their category labels (for example, handwritten digits or object classes). CNN hidden layers perform the operation of convolution between the input data and a set of filters, whose parameters are learned during training. The output of the convolution is further passed through a non-linear activation function. As the intermediate representations get passed through the multi-layered CNNs, the convolutions are performed over larger and larger image areas, and similar to biological vision systems the CNN layers are said to have increasing receptive fields. CNNs made a major breakthrough in vision, but when it comes to speech and language, temporal information and signal history play a very important role in defining the present and upcoming signal properties. *Recurrent neural networks* (RNNs) were designed to handle complex temporal data. Long short-term memory units (LSTMs) were developed to tackle technical challenges concerning vanishing gradients in RNNs (Hochreiter and Schmidhuber, 1997). LSTM hidden layers gate incoming information modulating what part of signal needs to be preserved, combined with the past, passed on further or entirely forgotten.

The optimization of ANN parameters is achieved through *back-propagation* of the cost function. In our case, the cost function is either mean least squares or softmax depending on the problem at hand. In this thesis we will also employ several optimization techniques for better training of ANNs, including dropout to tackle overfitting (Srivastava et al., 2014), batch normalization for improved performance (Ioffe and Szegedy, 2015) and weight initialization for faster convergence (Glorot and Bengio, 2010).

As mentioned above, the features extracted with ANNs received a lot of attention from the community due to their high interpretability and similarity to biological systems. In cognitive neuroscience, various features extracted from ANNs have been used to explain brain responses by fitting linear neural encoding models using ANN features as input (Cichy et al., 2016; Güçlü and van Gerven, 2015a, 2015b; Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016). For example, for ANN with two hidden layers, trained on image pixel data $\mathbf{X}$ to predict image category labels $\mathbf{l}$, there will be two sets of intermediate representations of input $\mathbf{X}$: $f^{(1)}(\mathbf{X})$ and $f^{(2)}\left(f^{(1)}(\mathbf{X})\right)$ for each hidden layer respectively. Then, each of these representations can be fitted to the neural responses in a separate linear neural encoding model in an attempt to explain the neural responses to images through the features of the image processing ANN:

$$\mathbf{l} = f^{(3)}\left(f^{(2)}\left(f^{(1)}(\mathbf{X})\right)\right) \tag{1.11}$$

$$\mathbf{y}_a{}^{(1)} = \boldsymbol{\beta}_a{}^{(1)\mathrm{T}} f^{(1)}(\mathbf{X}) + \boldsymbol{\varepsilon}_a{}^{(1)} \tag{1.12}$$

$$\mathbf{y}_a{}^{(2)} = \boldsymbol{\beta}_a{}^{(2)\mathrm{T}} f^{(2)}\left(f^{(1)}(\mathbf{X})\right) + \boldsymbol{\varepsilon}_a{}^{(2)} \tag{1.13}$$

Another approach to assess the relationship between the features of the ANN and the neural responses is RSA (see Data Exploration), that estimates the degree of similarity between the representations in the brain and the ANN. This has been done for various cognitive domains including vision (Güçlü and van Gerven, 2015a; Khaligh-Razavi and Kriegeskorte, 2014), auditory perception (Güçlü et al., 2016; Kell et al., 2018), memory (Hasson et al., 2015) and language (Jain and Huth, 2018). Many of these works underline the tight relationship between the features of the ANN and the brain (for review see Kietzmann et al., 2018; Kriegeskorte, 2015).

### *Nonlinear neural encoding models*

A tempting step forward from using features of ANNs trained on external stimuli (images, texts, sounds) is combining feature extraction with fitting onto brain responses in a single framework. This framework will then constitute a nonlinear neural encoding model. One would then train an ANN to directly predict neural responses from unprocessed stimulus data.

$$\mathbf{y}_a = \varphi(\mathbf{X}) \tag{1.14}$$

This approach, however, is associated with a number of challenges and therefore is not common in the field. The challenges include the limited amount of neural data (compared to available datasets of images and texts) and inherent noise in the neural responses (which might lead to incorrect input-label associations and prevalence of bad examples). Some previous work has been attempted in vision (Wen et al., 2018), however, the majority of the studies in computational neuroscience so far have used ANNs as feature extractors and fitted individual linear neural encoding models per each level of feature representation (each hidden layer of the ANN).

## Electrocorticography for studying language processing

There are many techniques that record neural data, ranging from single cell to whole-brain neural population recordings. The techniques are known to vary in the temporal and spatial resolution of the acquired neural signal and differ with respect to their signal-to-noise ratios and the underlying biological properties of the recorded signal.

In order for the data-driven approaches to yield best results (and particularly in case of deep ANNs) it is preferred to obtain high-quality, long-duration neural recordings. When studying language processing, high-quality recordings would mean neural data acquired at high temporal and spatial resolution and excellent signal-to-noise ratio.

Among the techniques with highest both temporal and spatial resolution is electrocorticography (ECoG). This neural acquisition modality samples neural responses subdurally, which means it records directly from the brain tissue is therefore only available in clinical setting. Some patients admitted with medicine-resistant epilepsy undergo ECoG recordings for localization of the source of epilepsy for its subsequent resection. For this, each patient undergoes an implantation surgery during which a grid of intracranial electrodes (each electrode of 1 cm or less in diameter) is placed on the patient's brain tissue. The grid then stays in place for approximately a week for monitoring and recording of the underlying brain activity. During this time, the patients can participate in research tasks.

The size and position of ECoG electrodes determine their spatial resolution. In general, ECoG data is known to provide exceptional signal-to-noise ratio and record responses of local neuronal populations of < 100,000 neurons under each electrode (Ritaccio et al., 2011). Temporal resolution it provides is typically > 250 Hz, which is more than necessary for studying various cognitive functions. One possible challenge of working with ECoG data is due to epilepsy, often accompanying structural and functional changes in the affected neural tissue. However, these changes are patient-specific and typically the majority of patients retain high-level cognitive processing exhibiting neural responses similar to the healthy controls (Ritaccio et al., 2010). Another challenge is associated with sparse spatial sampling of ECoG grids, where only activity directly underneath the electrode is being measured and activity of neural tissue between the electrodes is not recorded. In general, both these challenges can be tackled by combining data across multiple subjects to compensate for possible subject-specific cognitive deficits and improve density of cortical coverage.

ECoG has been used in a large number of studies investigating various aspects of cognition and language in particular (Crone et al., 2001; Ding et al., 2016; Flinker et al., 2011; Hermes et al., 2014; Lerner et al., 2011; Towle et al., 2008). Typically, researchers focus on the frequency-transformed ECoG signal (Ritaccio et al., 2013). The high-frequency band component, in particular, has been shown to reflect responses of the local neuronal populations driven by perception of external input. The pattern of the induced response in high-frequency band was shown to correlate with the firing rates of the underlying neuronal populations (Crone et al., 1998; Ray et al., 2008).

### Naturalistic experimental paradigms for studying naturalistic, continuous speech

Above we opted for electrocorticography to ensure top quality and high temporal and spatial resolution of the neural recordings. However, in order to maximize the benefit of applying data-driven approaches to these data we should consider what would be the optimal choice of the stimuli material and the experimental paradigm. Specifically, if we seek to discover language related structures in the brain data in a bottom-up way we should choose an experimental paradigm that incorporates the least amount of theoretical constraints and allows the participants to experience speech and language perception in an unconstrained naturalistic setting.

Simple tasks, such as listening to a story or watching a feature film, impose little cognitive demand on the participants, making it easier for them to attend to the stimuli and feel overall more interested and engaged in the experiment. Not only is this experimental setup better for the participants, it also allows for examining multiple language properties simultaneously and at multiple levels of hierarchical complexity, and lends itself to application of bottom-up approaches to extract structure in the collected recordings.

Naturalistic experimental paradigms have already been successfully employed to study neural mechanisms underlying complex cognition (Brennan, 2016; Hasson et al., 2010; Lerner et al., 2011). Successful previous work using naturalistic paradigms in combination with data exploration methods found strong similarity in how people process these complex stimuli and gained a lot of insight about the details of neural processing (Hasson et al., 2004; Lerner et al., 2011).

Importantly, relaxing constraints on cognitive demands not only leads to better engagement of participants in the experiment, it also facilitates long-duration recordings of brain activity. Long-duration recordings result in large amounts of data, which is one of the factors determining success of the data-driven approaches aimed at complex feature extraction for fitting onto the neural responses.

Thus, naturalistic experimental paradigms provide a flexible data collection framework that produces rich multi-purpose neural and behavioral datasets. Here we opt for a naturalistic experimental paradigm, such as a film-watching experiment because it combines all the aforementioned benefits of naturalistic stimulation. Experiments using films reported high levels of attention and engagement among subjects. This factor is particularly relevant in case of ECoG where patients often find it difficult to concentrate on tasks of high cognitive demand for long periods of time. Combination of both visual and auditory modalities not only makes the task more engaging compared to listening or reading of a story, it also opens up new possibilities for analyses of the brain responses. Presence of both perceptual modalities poses challenges caused by the interaction of sensory channels, but at the same time, it makes the collected data even richer, allowing for investigation of the representations that arise from both auditory and visual input as well as studying their interaction effects.

## 1.2 OUTLINE

The present work focuses on application of data-driven approaches to the ECoG neural responses collected during speech and language processing in a context of a simple naturalistic task, such as perception of an audiovisual film. Specifically, we focus on neural data exploration and development of high-performing neural encoding models that explain the brain responses during speech perception and provide an interpretation for the functional involvement of the responsive cortex.

Most of the projects in the present thesis are carried out using data from clinical ECoG grids. However, we also collected and analyzed high-density ECoG, which provides an even finer detail of neural recordings due to a small size of individual electrodes (≈1 mm in diameter) and high density of coverage (128 channels over 3-4 cm).

Overall, we used two film-watching experiments: a short extract from a feature film (6.5 minutes) and a full-length feature film (78 minutes). Depending on the project goals, we either use one dataset or a combination of both.

In the following chapters, we use different techniques to extract and map various speech and language properties onto the brain responses. Since we generally adhere to bottom-up methods, our overall approach is to use low-level information as input to a model predicting the brain responses. In the process of this mapping the low-level features may go through various transformations to optimize the fit. Various data exploration techniques are used to interpret the relationship between the well-fitted neural sources and the features used.

Thus, in **Chapter 2**, we start with encoding of low-level acoustic speech features in the ECoG neural responses. We train a linear neural encoding model on the acoustic properties of speech to predict responses along the perisylvian cortex. The results of this work provide a baseline in a linear mapping of a predefined set of low-level speech features onto the ECoG responses. In a way, the low-level features used here can be considered theory-driven, as their definition and configuration are predetermined by a theoretical framework that models acoustic processing by the cochlea and the primary auditory cortex. The features are essentially the Gabor-filtered spectrogram properties that detect temporal and spectral changes in sound, similar to the edge detectors in visual processing. We use a clustering approach to investigate tuning of individual cortical sites to different groups of features. Labels of various linguistic units (from phonemes to words) are used post-hoc to interpret the results of this tuning.

In **Chapter 3**, we expand on the approach from Chapter 2. Here, rather than using a predefined set of speech features we develop a bottom-up artificial neural network (ANN) model for automatic extraction of sound and speech features that drive the neural responses. Thus, in this project we train a non-linear neural encoding model on unprocessed sound input to extract optimal sound and speech features that would best fit the brain data. We aim to determine whether such a model can be successful and general enough to fit the neural responses not only in a held-out test set but across subjects and for different stimulus materials. Thus, the model is trained on the neural responses to the long-duration film and is further validated on the data from the short feature film and in a different set of subjects. Apart from gaining in accuracy of predicting the brain data, we are interested in analyzing and interpreting the bottom-up features that the model extracts. In particular, we are curious if the observed results agree with the current theoretical frameworks of speech and auditory processing and whether the collected data-driven evidence can be used for expanding our knowledge and generating new hypotheses of speech perception in the human brain.

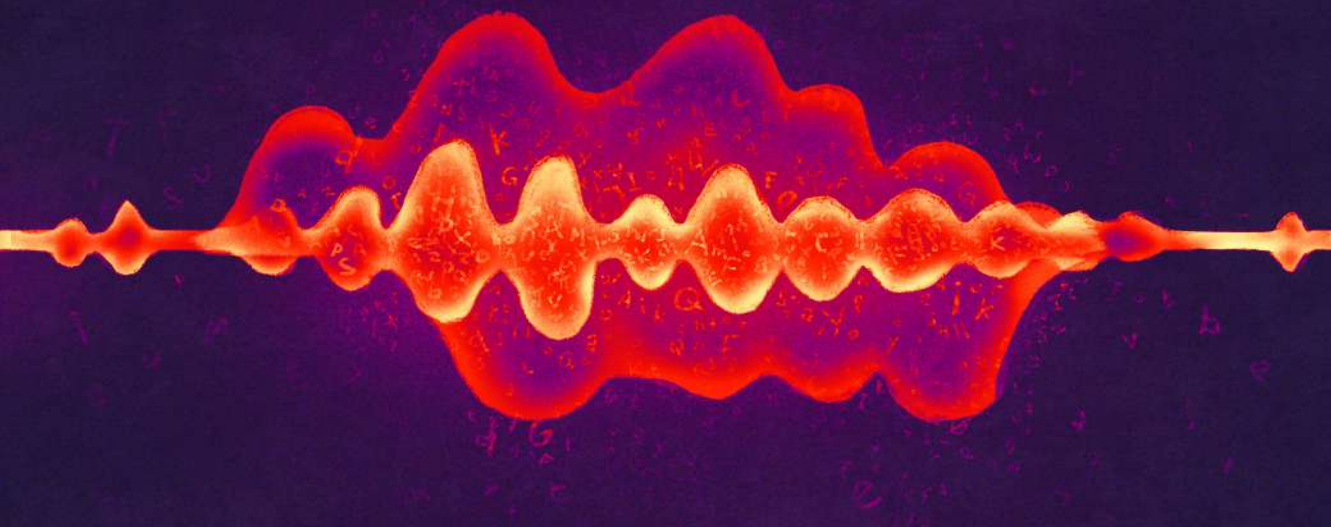
In **Chapter 4**, we move beyond the perisylvian and auditory cortex and investigate the ECoG neural responses to speech in the sensorimotor cortex. Specifically, we apply exploratory analyses to high-density ECoG data in an attempt to find patterns triggered by perception of speech. We use quite a generic speech property, such as spectral sound envelope, to investigate if the neuronal populations in pre- and postcentral cortices respond to it. Due to the fact that the primary function of this cortex is sensorimotor processing we also aim to interpret the observed result in the context of motor functions including hand movements and speech articulation.

In **Chapter 5**, we go beyond the auditory perception and engage the visual stream of the film. In particular, we develop a pipeline that processes low-level visual input of the film and extracts from them more high-level semantic information that can be used to predict the neural responses across multiple cortical regions. Thus, once again we use a bottom-up ANN based approach to extract high-level semantic features from the low-level visual properties of the feature film and use them to explain the variance in the associated ECoG responses. First, the low-level visual features are passed through an automatic object recognition ANN that identifies visual objects and concepts present in the film frames. This information is then passed through a language ANN that combines the visual object data per each frame and transforms it into a high-level semantic representation. These representations are then fit onto the brain responses to test the similarity of the semantic concepts extracted using bottom-up ANN methods and the cortical processing mechanisms.

chapter

2

# Neural tuning to low-level features of speech throughout the perisylvian cortex



# CHAPTER 2. NEURAL TUNING TO LOW-LEVEL FEATURES OF SPEECH THROUGHOUT THE PERISYLVIAN CORTEX1

Despite a large body of research, we continue to lack a detailed account of how auditory processing of continuous speech unfolds in the human brain. Previous research showed the propagation of low-level acoustic features of speech from posterior superior temporal gyrus towards anterior superior temporal gyrus in the human brain (Hullett et al., 2016). In this study we investigate what happens to these neural representations past the superior temporal gyrus, and how they engage higher-level language processing areas, such as inferior frontal gyrus. We used low-level sound features to model neural responses to speech outside the primary auditory cortex. Two complementary imaging techniques were used with human participants (both males and females): electrocorticography (ECoG) and functional magnetic resonance imaging. Both imaging techniques showed tuning of the perisylvian cortex to low-level speech features. With ECoG, we found evidence of propagation of the temporal features of speech sounds along the ventral pathway of language processing in the brain towards inferior frontal gyrus. Increasingly coarse temporal features of speech spreading from posterior superior temporal cortex towards inferior frontal gyrus were associated with linguistic features, such as voice onset time, duration of the formant transitions, as well as phoneme, syllable and word boundaries. The present findings provide groundwork for a comprehensive bottom-up account of speech comprehension in the human brain.

[1] This chapter is based on Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A., & Ramsey, N. F. (2017). Neural tuning to low-level features of speech throughout the perisylvian cortex. *Journal of Neuroscience*, 37(33), 7906-7920.

## 2.1 INTRODUCTION

Speech is a specific kind of complex sound, and, similarly to complex object processing in vision (for example, face processing), a substantial amount of research has focused on the neural tuning to low-level properties of this type of sensory input. Comparing tones revealed that primary auditory cortex (PAC) and immediately adjacent superior temporal gyrus (STG) are tonotopically organized (Wessinger et al., 1997; Bilecen et al., 1998; Formisano et al., 2003) and modulated by various spectral and temporal properties of sound (Giraud et al., 2000; Joris et al., 2004; Altmann et al., 2010; Leaver and Rauschecker, 2010).

There is also a large body of literature that describes speech in the context of language theory. These studies relate different higher-level language features, from phonemes to phrase units, to the neural responses to speech throughout the cortex (Giraud and Poeppel, 2012; Hagoort and Indefrey, 2014; Ding et al., 2015). Altogether, these studies outline a network of key language processing regions: STG, anterior temporal cortex, inferior frontal gyrus (IFG), middle temporal, precentral and angular gyri (Hickok and Poeppel, 2007; Friederici, 2012; Hagoort, 2013). However, establishing a connection between low-level auditory features and higher-level language structures in continuous speech remains a challenging task.

In the present study, we focus on the neural tuning to acoustic features of speech along the perisylvian cortex. We believe that understanding the anatomical and temporal characteristics of the neural tuning to the low-level features of naturalistic speech will provide groundwork for a bottom-up account of speech perception. Hopefully, a comprehensive bottom-up account of speech perception can be further integrated into existing theoretical frameworks of language processing in the brain (Hickok and Poeppel, 2007; Friederici, 2012; Hagoort, 2013), altogether providing an extensive model of neural processing of speech.

Here we used continuous naturalistic stimuli, which were segments from a feature film (Pippi Longstocking, 1969), to investigate the neural tuning to speech and compare it to the neural tuning to another type of complex sound – music. We used linear regression to model how neural responses are elicited by low-level sound features. The neural responses to the movie were obtained using electrocorticography (ECoG). Subsequently, we obtained and processed analogous functional magnetic resonance imaging (fMRI) data to assess the reproducibility of ECoG findings with a more commonly accessible technique and at higher spatial sampling.

Our results suggest a propagation of temporal features of the speech audio signal extending beyond the auditory network into the IFG during speech comprehension. The temporal features passed along the pathway were associated with distinct linguistic markers (from subphonemic features to syllable and word boundaries). The present evidence suggests that, like in the visual domain, increasingly complex stimulus features are processed in downstream areas, pinpointing the similarity in neural coding across perceptual modalities.

## 2.2 RESULTS

Segments from a feature film Pippi Longstocking (1969) were edited together to make a 6.5-minute video with a coherent plot. The video was accompanied by a soundtrack which contained contrasting fragments of speech and music, 30 seconds each. This naturalistic stimulus was used in both ECoG and fMRI experiments of the present study. During both experiments the participants were asked to attend to the short movie. The linear model using low-level sound features was trained on ECoG and fMRI data, separately. Due to its high spatial sampling, fMRI results were used as complementary evidence of the cortical topography of model performance and feature tuning, observed in ECoG.

We first explored the low-level acoustic properties of the soundtrack of the movie. The NSL toolbox (Chi et al., 2005) was used to obtain spectrotemporal modulations (SMs and TMs) of the sound based on the audio spectrogram (**Figure 2.1a**). These modulations can intuitively be interpreted as representations of sound at different resolutions along the spectral and temporal dimensions. Finer modulations capture the sound at a higher resolution and thus reflect finer changes along the spectral and temporal dimensions. Coarser modulations capture the sound at a lower resolution and reflect smoothed energy distributions along both dimensions.

In the soundtrack of the movie used in the present study, low-level spectrotemporal features differed between speech and music fragments, as assessed with two-tail *t*-tests, run for each SM-TM feature (p < .001, Bonferroni corrected). Speech exhibited a larger spread along the TM dimension (2 – 8 Hz) than music (< 2 Hz). Conversely, music exhibited a larger spread over the SM dimension (1 – 16 cyc/oct at TMs < 2 Hz), compared to speech (< 4 cyc/oct) (**Figure 2.1b** and **c**).

### Neural tuning to low-level features of speech and music in ECoG: model performance and topography

#### Encoding model performance

Fifteen patients watched the movie in an ECoG experiment. Previous studies have shown that the high frequency band (HFB) amplitude closely correlates with neuronal firing rates and with fMRI blood-oxygenation-level-dependent response (Hermes et al., 2012; Lachaux et al., 2007). Here we extracted the 60 – 120 Hz amplitude from the signal. A linear kernel ridge regression model was applied to predict neural responses from the sets of SMs (0.03 – 8 cyc/oct) and TMs (0.25 – 64 Hz). To account for time differences between neural responses and sound onset, the analysis included time lags between 0 and -500 ms, which constituted a third feature dimension. Per electrode, the model was trained on

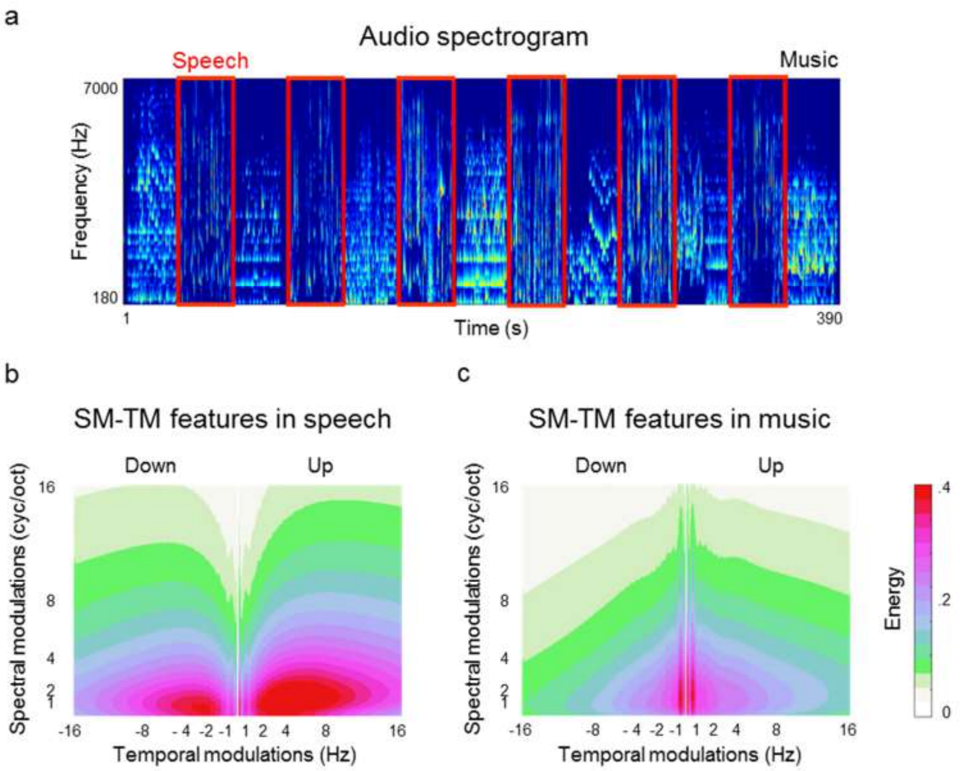


**Figure 2.1. Low-level properties of sound. (a)** Audio spectrogram of the soundtrack of the 6.5 min video stimulus. Horizontal axis represents time and vertical axis represents frequency scale. The spectrogram was obtained using NSL toolbox at 8 ms windows (sampling rate = 125 Hz). The frequency bins are spaced logarithmically (128 bins) in the range of 180–7000 Hz. Speech blocks are framed in red. Music blocks are the rest. **(b)** Low-level spectrotemporal (SM–TM) features of speech. Horizontal axis represents TMs measured in Hertz; vertical axis represents SMs measured in cycles per octave (cyc/oct). The features were extracted using the NSL toolbox by convolving the spectrogram with a bank of 2D Gabor filters at different spatial frequencies and orientations. The SM–TM features were extracted for every time point and every frequency bin. TMs were obtained at 0.25–16 Hz, in linear steps of 0.2 Hz. SMs were obtained at 0.25–16 cyc/oct in linear steps of 0.2 cyc/oct. Positive and negative sign representations along the TM axis capture the direction of energy sweeps in the spectrogram (up and down). The extracted SM–TM features were averaged over time points within block and across blocks. **(c)** Low-level SM–TM features of music. The feature extraction and visual representation are identical to the SM–TM features of speech in (b).

speech and music data combined, and was tested to predict HFB responses to speech and music separately. All individual accuracy scores were thresholded at p < .001, Bonferroni corrected. The model predicted the neural responses evoked by speech in 10% of all electrodes (**Figure 2.2a**). The neural responses evoked by music could be predicted in 3%

of all electrodes (**Figure 2.2b**). The accuracy scores varied considerably depending on the cortical region, but on average, across patients, the maximal accuracy score ($r_{max}$) reached .39 ± .08 when predicting neural responses to speech, and .18 ± .05 when predicting neural responses to music. In general, modeling neural responses to speech and music produced high prediction accuracy for non-overlapping sets of electrodes. Examples of observed and predicted neural responses are shown in **Figure 2.2c** for speech and **Figure 2.2d** for music.

### Low-level speech encoding: a posterior-anterior prediction accuracy gradient

In the case of speech, we observed a posterior-anterior prediction accuracy gradient along the perisylvian cortex: prediction accuracy values were highest in posterior STG, and declined as the location of the electrode moved towards IFG (**Figure 2.2a**). The effect was significant as tested with a one-way ANOVA comparing anatomical electrode location to prediction accuracy (F(4) = 3.25, p = .02). Locations were confined to five cortical regions covering the anterior-posterior axis along the perisylvian cortex: IFG pars triangularis and pars opercularis, precentral gyrus, postcentral gyrus and STG. Electrodes in STG exhibited higher accuracy compared to electrodes in pars opercularis and pars triangularis: $p = 5 \times 10^{-3}$ and p = .01, respectively, based on least significant difference test. The ANOVA test showed a difference in model performance between the cortical regions. To quantify the effect of the poster-anterior direction of the difference in model performance, we assigned numerical labels to the five cortical regions based on the ranked distance from STG (1 to 5) and fit a linear trend on the prediction accuracy along the five regions. The model fit was significant compared to the constant model: $\beta_{label} = 0.02$, F(100) = 12.9, $p = 5.1 \times 10^{-4}$ and $\beta_{intercept} = 0.16$ ($p = 3.5 \times 10^{-8}$).

### Distinct tuning profiles across cortical sites in ECoG

Next, we explored whether different cortical sites, where responses were well predicted by the model, exhibited distinct feature tuning profiles. Examples of feature tuning profiles are shown in **Figure 2.2e** for speech and **Figure 2.2f** for music. We aimed at uncovering brain sites with similar feature tuning in an unsupervised fashion to avoid superimposing any assumptions about the topography of the tuning profiles. Thus, an affinity propagation clustering (Frey and Dueck, 2007) analysis was performed on the regression coefficients, z-scored over the electrodes. The electrodes with similar regression coefficients were clustered together as having a similar feature tuning profile. The clustering analyses were performed separately for speech and music, as modeling neural responses to speech and music resulted in different sets of electrodes. Permutation testing was used to assess the ranges over the three feature dimensions (TMs, SMs and time lags) that the clusters were significantly tuned to.

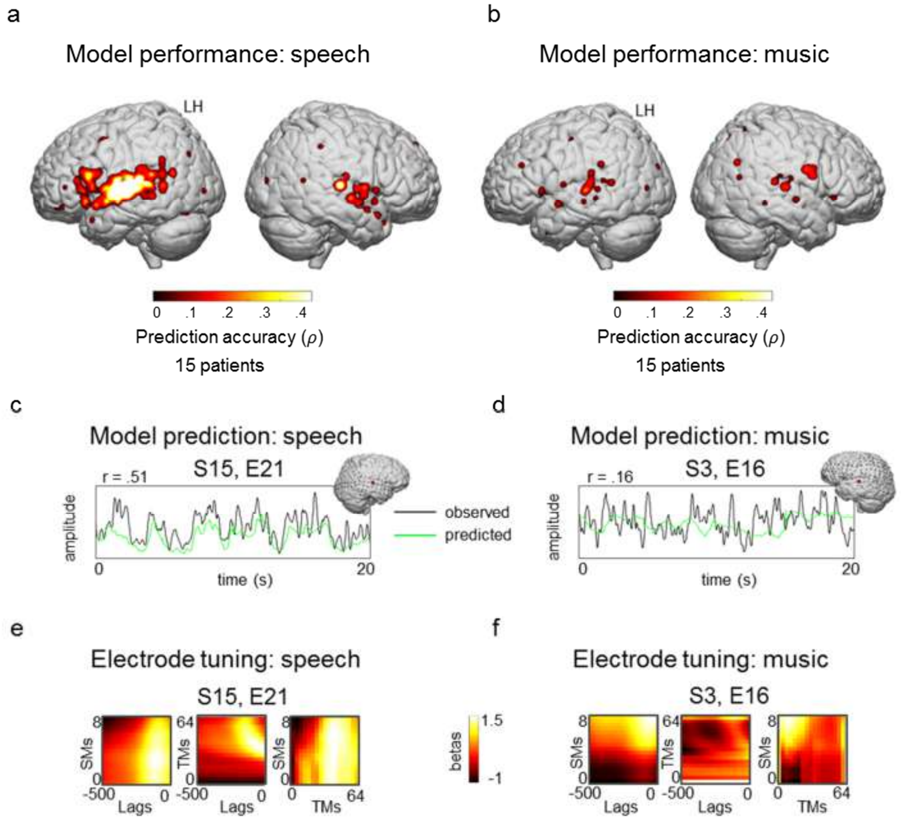


**Figure 2.2 Encoding model performance in ECoG.** A linear encoding model was trained on low-level SM–TM features to model neural responses in ECoG evoked by speech and music. The model was trained per individual electrode. The model performance was quantified in terms of prediction accuracy: the held-out neural responses (test set) were correlated to the corresponding time courses predicted by the model trained on low-level SM–TM features of the audio. The model performance assessment was performed separately for speech and music. Spearman correlation scores were used. The presented maps were thresholded at p < 0.001, Bonferroni corrected for the number of electrodes (1283 electrodes over 15 patients). For the visualization purposes, a 2D Gaussian kernel (FWHM = 10 mm) was applied to the coordinate on the brain surface corresponding to the center of the electrode, so that the prediction accuracy value obtained by the model faded out from the center of the electrode toward its borders. The results are shown on a standard MNI brain. Individual electrode locations were normalized to the MNI space using patient-specific affine transformation matrices obtained with SPM8. **(a)** Prediction accuracy maps for modeling neural responses to speech (130 electrodes). **(b)** Prediction accuracy maps for modeling neural responses to music (47 electrodes). **(c)** Example of a neural response predicted by the model (green) plotted against the observed response (black) in speech. **(d)** Example of predicted and observed neural responses in music. **(e)** and **(f)** are examples of electrode feature tuning profiles: (e) shows electrode from (c) and (f) shows electrode from (d). Non z-scored regression coefficients were used.

## Neural tuning to temporal and spectral characteristics of speech

In the case of speech, the affinity propagation clustering returned six clusters (Clusters 1 – 6), each of which contained electrodes from about half of all patients. Across clusters most variance was concentrated along the dimensions of TMs and time lags (**Figure 2.3**). Electrodes in Cluster 2 exhibited tuning to fast TMs (11 – 45 Hz) within a 150 ms window after the sound onset. Conversely, electrodes in Cluster 6 were tuned to slow TMs (0.7 – 4 Hz) within a 315 – 500 ms window after the sound onset. Clusters 1 – 6 did not exhibit distinct SM tuning, but showed some preference to coarse SMs (< 1 cyc/oct). From the anatomical point of view, Clusters 1 – 3 comprised electrodes primarily in posterior STG, whereas Clusters 4 – 6 comprised electrodes in IFG and anterior STG. Clusters 2 – 5 comprised a small number of electrodes in supramarginal gyrus.

With respect to the negative regression coefficients (**Figure 2.3**), some electrodes showed negative tuning to certain features prior to any z-scoring procedures. We believe the negative tuning indicates a decrease in the HFB amplitude of an electrode relative to its mean HFB amplitude as a response to the stimulus (estimated by the bias term).

Additionally, we looked at the average tuning profiles per anatomical region along the perisylvian cortex rather than the clusters. The tuning plots per anatomical region corresponded to the ones recovered with the clustering (**Figure 2.3**), but exhibited more variance in tuning along each feature dimension. Moreover, averaging the regression coefficients over the whole STG did not result in any specific tuning, whereas using the clustering approach allowed us to uncover a number of functional subdivisions along STG (Clusters 1 - 5).

No hemisphere specific tuning profile was found across the patients. In general, the electrodes located similarly across the hemispheres were assigned to the same cluster.

Noticeably, Clusters 2 – 3, as well as Clusters 4 – 5 showed similar cortical topography with varying tuning profiles. These local variations could be due to the anatomical variability among the patients. Overall, the change in tuning to TMs and time lags followed a global trend: from tuning to shorter lags and faster TMs in posterior STG towards tuning to larger lags and slower TMs in IFG and anterior STG. The change in tuning to TMs and time lags mapped on the posterior-anterior prediction accuracy gradient as estimated with a stepwise linear regression: $\beta_{lag} = 1.33 \times 10^{-4}$, $\beta_{TM} = 9.67 \times 10^{-4}$, F(127) = 13.3, p = $5.92 \times 10^{-6}$.

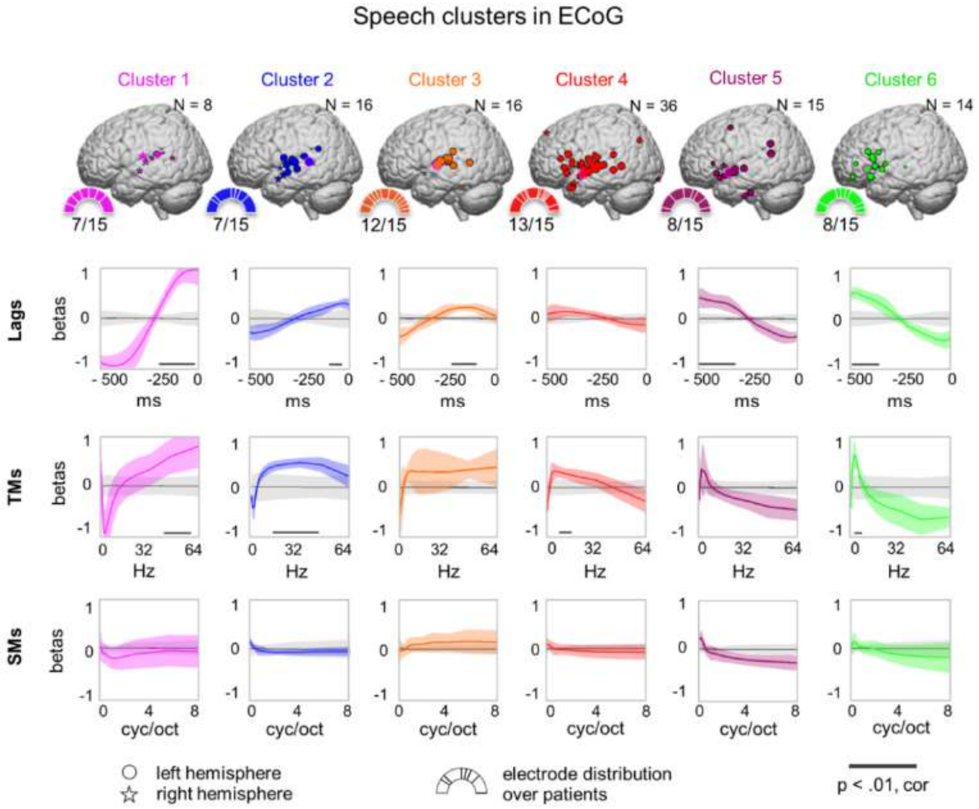


**Figure 2.3. Affinity propagation clustering of regression coefficients in ECoG electrodes tuned to speech.** Shown are six clusters produced by clustering on regression coefficients of electrodes involved in speech tuning. For each cluster, a number of properties are shown: the number of electrodes belonging to the cluster (*N*), the normalized electrode locations on the MNI brain, the distribution of electrodes over patients (half pie charts with number of patients out of 15), and the feature tuning profiles over lags, TMs, and SMs. Right hemisphere electrodes were projected on the left hemisphere and marked with star symbols; circles show left hemisphere electrodes. The size of electrodes on the MNI brain reflects the similarity of the electrode's feature tuning profile to the exemplar electrode of the cluster (framed in pink). The sectors of the half pie charts show percentage of each patient's electrodes relative to the overall number of electrodes in the cluster (*N*). The lags were used in the range of 0 to −500 ms with respect to audio onset; TMs were in the range of 0.25–64 Hz; SMs were in the range of 0.03–8 cyc/oct. Per cluster, mean (colored bold curve) and SD (colored shading) over regression coefficients are plotted for each dimension. The mean (gray bold curve) and SD (gray shading) of the null distribution based on permutations of cluster assignments are shown. The black thick line at the bottom of several plots represents significant segments of the tuning profile compared with the null distribution ($p < 0.01$, Bonferroni corrected).

We further explored the effect of time lag on the model performance across different clusters by retraining the model with reduced lag feature sets and comparing the model performances. In addition to using audio features at all lags from 0 to -500 ms (*Lag0-500*), we retrained the model using no input lag information (*Lag0*) or using only audio features at a lag of -500 ms (*Lag500*). Interaction between lag information and assignment to Clusters 1 – 6 was tested using a two-way ANOVA with unbalanced design: F(10) = 10.32, p = 6.06 × $10^{-15}$ (**Figure 2.4**). Prediction accuracy in Cluster 1 was higher for *Lag0* and *Lag0-500* compared to *Lag500*: p = 4.64 × $10^{-6}$ and p = 5.32 × $10^{-6}$, respectively. Prediction accuracy in Cluster 6 was higher for *Lag500* and *Lag0-500* compared to *Lag0*: p = 1.7 × $10^{-3}$ and p = 3.7×$10^{-3}$, respectively. In the case of *Lag0* more electrodes were modeled in Clusters 1 – 3, compared to *Lag500*; the opposite was observed for Clusters 4 – 6. Despite the lag difference, in all cases the encoding model learned the same tuning to SMs and TMs in all clusters as quantified with multiple one-way ANOVA tests on cluster mean regression coefficients, separately for TMs and SMs (TMs: 0.1 ≤ F(2) ≤ 0.52, SMs: 0.77 ≤ F(2) ≤ 4.27, both at p > .01).

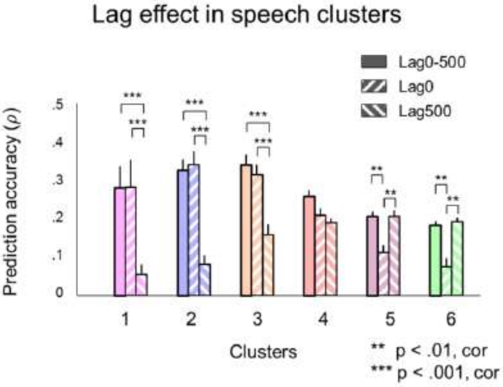


**Figure 2.4. Lag effects in encoding of speech in ECoG responses.** In addition to the full model using all lags from 0 to –500 ms of the audio features to predict ECoG responses (Lag0–500), two more models with reduced lag representation were trained. The first additional model included no lags at all (Lag0) and the second additional model only included one lag at –500 ms (Lag500). The bars show mean prediction accuracy over all electrodes in the cluster. The error bars indicate SEM. A two-way ANOVA test with unbalanced design was performed to estimate the main effect of the lag factor and the interaction between the lag factor and cluster assignment (on 105 electrodes). Groups with significantly different means are shown at $p$ < 0.01 and $p$ < 0.001. The $p$-values were corrected for multiple comparisons using Fisher's least significance difference under the condition that the null hypothesis had been rejected: $F_{(10)}$ = 10.32, $p$ = 6.06 × $10^{-15}$.

## Link between low-level TMs and rates of language features

In order to obtain a more intuitive interpretation of the TM tuning results, we compared the TM tuning profiles of the different clusters with the temporal rates of well-known language features. We estimated the temporal rates (see Methods) for a number of critical language features from the acoustic data (**Figure 2.5a**). We considered the rates of three features that represent important building blocks of language: phonemes, syllables and words. We also considered two features at the scale smaller than a phoneme: voice onset time and formant transitions. These two features are crucial for the categorization of a large number of phonemes and are therefore central to speech comprehension (Ganong, 1980; Liberman et al., 1967; Mesgarani et al., 2014). Voice onset time is a characteristic of plosive consonants and represents the amount of time between the release and the voice onset of the subsequent vowel. Based on this feature, voiced plosives, associated with shorter voice onset time, can be distinguished from voiceless plosives, associated with longer voice onset time. Formant transitions are characteristics of consonant-vowel combinations and are indicative of the place of articulation, and thus the category, of the neighboring phonemes. For all these features we calculated their temporal rates to see whether there might be a relationship between the language features and low-level TM tuning in the observed ECoG responses.

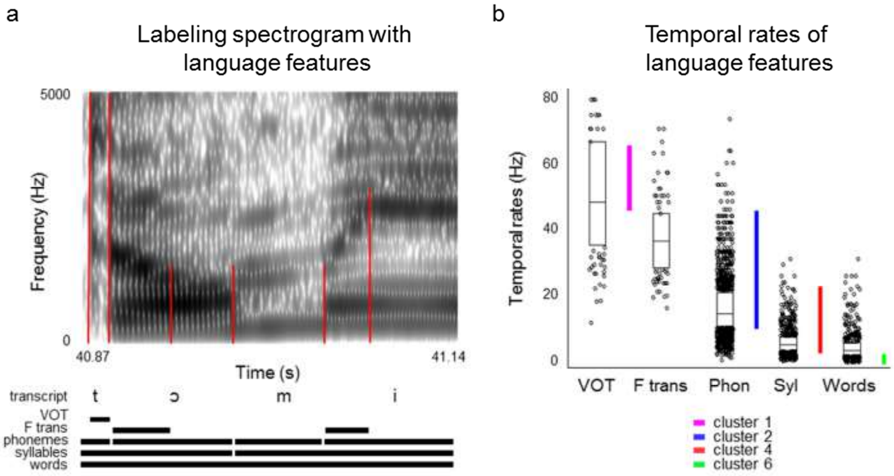


**Figure 2.5. Rates of various language features. (a)** Details of annotating the spectrogram with the language features. Based on the spectrogram, we identified the boundaries for formant transitions, voice onset time, phonemes, syllables, and words. Red lines outline the acoustic features based on which the annotations were made; for example, here: noise before the voicing of the vowel, changes in the second formant of the vowel, and boundaries between different formant structures of the vowels and a sonar consonant. **(b)** Distributions of rates for each language unit in Hertz: VOT (voice onset time), F trans (formant transitions), Phon (phonemes), Syl (syllables), and Words. The rates

were calculated based on the duration of each instance of each language feature. The durations were then expressed in Hertz. The lower and upper boundaries of the box plots show 25th and 75th quantile, respectively. The horizontal bars show the median. The dots are the remaining rate values. Colored vertical bars show TM tuning profiles for clusters 2, 4, and 6.

On average, the language markers of phoneme and syllable boundaries showed moderate correspondence to TM tuning of Cluster 2 and Cluster 4, respectively (**Figure 2.5b**). Subphonemic features, such as voice onset time and formant transitions changed at a temporal rate similar to the TM tuning of Clusters 1 and 2 (**Figure 2.5b**). The TM tuning profile of Cluster 6 seemed to overlap to some extent with the temporal rates of word boundaries. However, it appeared that our naturalistic speech stimuli contained a lot of monosyllabic words, and the temporal rate of word boundaries was similar to the temporal rate of syllable boundaries. Cluster 6 showed tuning to potentially slower TMs than the rate of individual word transitions.

### Neural tuning to temporal and spectral characteristics of music

Analogous affinity propagation clustering of regression coefficients was performed for music encoding. Two clusters with opposite tuning profiles were revealed (**Figure 2.6**). Distinct SM and TM temporal profiles were uncovered, however, no specific time lag tuning was found. Cluster 3 included electrodes without specific tuning to any of the low-level features (**Figure 2.6**). No hemisphere specific tuning profile was found across the patients.

We tested to what extent the apparent difference in model performance between speech and music could be attributed to our choice in the range of low-level spectrotemporal features. Given that music exhibited a larger spread over the dimension of SMs, the model was retrained to predict neural responses to music on another set of SM features, spanning from 0.5 to 32 cyc/oct, keeping the TMs fixed (0.25 to 64 Hz). The resulting prediction accuracy scores did not differ significantly from the prediction accuracy scores obtained with the previously used set of features, based on one-tailed t-test: $t = 0.74$, $p = .46$. Thus, the limited range of SMs (0.03 – 8 cyc/oct) chosen for the encoding analyses, did not explain why neural encoding of music was less prominent than that of speech.

## Complementary findings with fMRI

Seven of the 15 patients, whose ECoG results are described above, watched the same movie in the MRI scanner. Analogous analyses were performed on the obtained fMRI data.

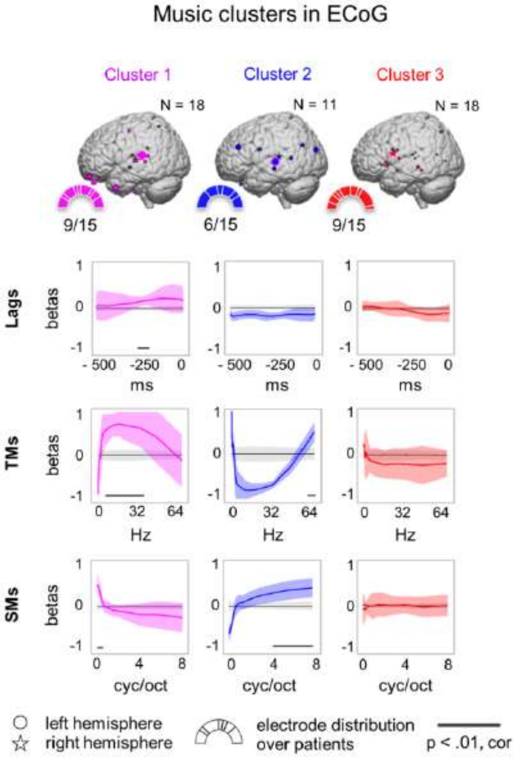


**Figure 2.6. Affinity propagation clustering of regression coefficients in ECoG electrodes tuned to music.** Clustering setup, significance testing, and the display format are identical to Figure 1.5. However, in the case of music encoding, a different set of electrodes was significantly well modeled based on low-level sound features (47 electrodes). The choice of cluster colors is based on convenience; there is no relationship between speech and music clusters represented by the same color.

A linear kernel ridge regression model was trained to predict voxel responses from SMs, TMs and audio frequency. Time lag was not included due to the slow temporal resolution of fMRI. Across patients, responses in STG were modeled best. On average across patients, the maximal accuracy ($r_{max}$) reached .54 ± .1 for speech encoding, and $r_{max}$ = .48 ± .08 for music encoding (all at p < .05, FDR-corrected). Distinct feature tuning profiles were estimated using affinity propagation clustering. The clustering was performed separately for speech and music, but in both cases two clusters were revealed. Cluster 1 comprised voxels throughout STG tuned to a broad range of TMs (3 – 45 Hz), coarse SMs (< 0.25 cyc/oct), and high frequencies (> 3 kHz) (**Figure 2.7b**). Cluster 2 comprised voxels in prefrontal, parietal and latero-occipital cortices, and did not exhibit distinct tuning profiles along any of the dimensions.

Finally, the anatomic similarity between ECoG and fMRI speech encoding results was quantified. For each individual ECoG electrode a corresponding sphere of voxels was defined within a radius of 10 mm around the electrode center coordinate. Five out of seven subjects showed significant correlation between prediction accuracy in ECoG electrodes and corresponding fMRI spheres: $r_1 = .35$ ($p = 4 \times 10^{-3}$), $r_3 = .44$ ($p = 5 \times 10^{-4}$), $r_5 = .35$ ($p = .02$), $r_6 = .53$ ($p = 3 \times 10^{-4}$), $r_7 = .36$ ($p = 5 \times 10^{-3}$). **Figure 2.7a** shows the overlap in model performance when using ECoG and fMRI data for one representative patient. Most similarity in prediction accuracy between ECoG and fMRI results was observed along STG.

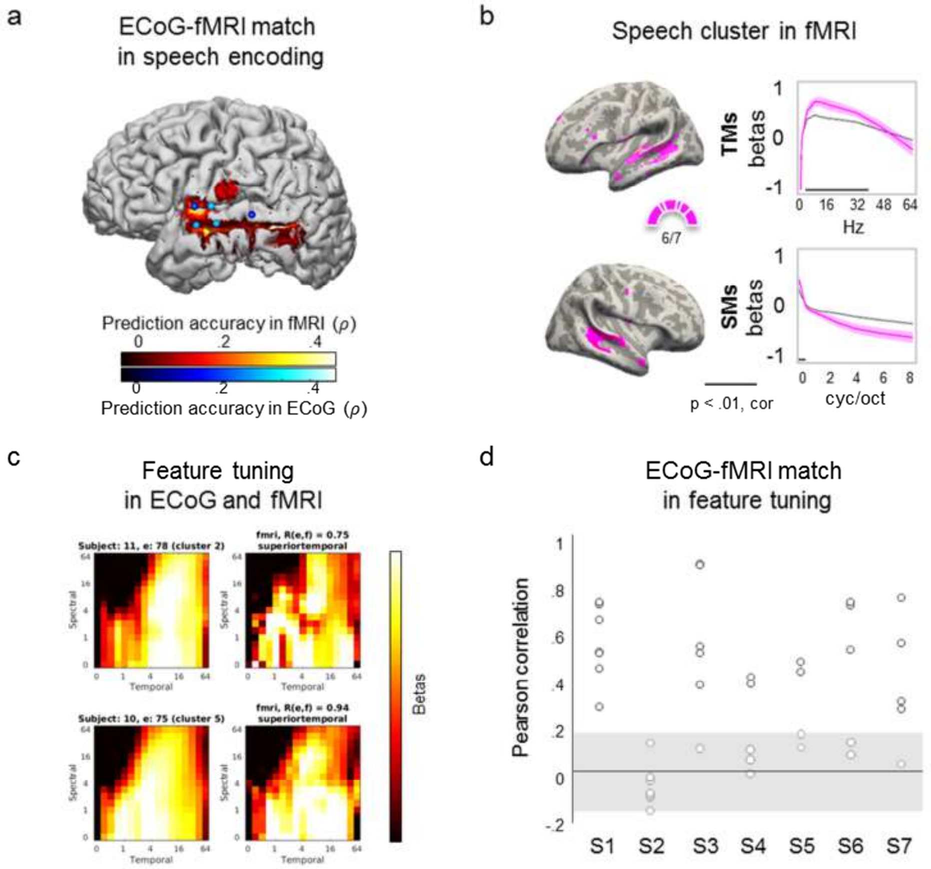


**Figure 2.7. Agreement between ECoG and fMRI in tuning to low-level features of speech. (a)** Prediction accuracy for modeling neural responses to speech with ECoG and fMRI for one representative patient. The map of electrodes tuned to low-level speech features (p < .001, Bonferroni corrected) in overlaid on top of the map of the voxels tuned to low-level speech features (p <.05, FDR corrected). For display purposes, the fMRI map was smoothed at FWHM = 10 mm. The results are shown in the native space of one patient's

brain. **(b)** A speech cluster in fMRI, matching ECoG speech Clusters 1 – 6. A similar affinity propagation clustering analysis was performed on the regression coefficients learned from the fMRI data. The black line at the bottom of the plots shows significant segments of the tuning profile, compared to the null distribution ($p < .01$, Bonferroni corrected). The displayed cluster showed a feature tuning profile along the perisylvian cortex, similar to speech Clusters 1 – 6 in ECoG. The fMRI speech cluster was tuned to fast TMs and coarse SMs. **(c)** Examples of SM-TM feature tuning profiles observed in ECoG electrodes and corresponding fMRI spheres. In each patient, for every ECoG electrode a corresponding fMRI sphere was calculated within a radius of 10 mm. The SM-TM feature tuning profile of the best modelled voxel within the sphere was correlated with the feature tuning profile of the corresponding electrode. Pearson correlation between the ECoG and fMRI-based SM-TM profiles is reported per example. **(d)** Pearson correlation between SM-TM feature tuning profiles in ECoG electrodes and corresponding fMRI spheres per patient. Only electrodes and voxels with significant model performance in speech were used. The null distribution based on 10000 permutation tests in shown in grey (0.1 and 99.9 percentiles). Significant correlation ($p < .001$, Bonferroni corrected) have black outlines, insignificant correlations are greyed out. Per patient, we report the number of electrodes with significant correlations compared to the number of electrodes used in the analysis: $N_1$= 7 (out of 7), $N_3$= 5 (out of 6), $N_4$= 2 (out of 5), $N_5$= 2 (out of 4), $N_6$= 3 (out of 5), $N_7$= 4 (out of 5).

Additionally, we assessed the similarity of feature tuning profiles in ECoG electrodes and corresponding fMRI spheres. We first considered speech encoding results. The SM-TM tuning profiles uncovered with ECoG and fMRI were correlated (**Figure 2.7c** and **d**). In six out of seven patients ECoG-based and fMRI-based feature tuning profiles were significantly correlated (**Figure 2.7d**): mean $r_1 = .57$ (100% of electrodes), mean $r_3 = .56$ (80% of electrodes), mean $r_4 = .16$ (40% of electrodes), mean $r_5 = .23$ (50% of electrodes), mean $r_6 = .41$ (60% of electrodes), mean $r_7 = .38$ (80% of electrodes). **Figure 2.7c** shows some examples of similarity of tuning profiles in ECoG electrodes and corresponding fMRI spheres. The overall contour of the feature tuning profile is reproduced in fMRI results, however the gradient of the TM tuning across the perisylvian cortex is not observed in the fMRI data. The latter is likely due to low temporal resolution of fMRI combined with the use of continuous, naturalistic speech stimuli. Most similarity in feature tuning profiles between ECoG and fMRI results was observed along STG. Analogous analyses were performed on the results of music encoding. However, for music, less than half of patients showed agreement between ECoG and fMRI.

## 2.3 DISCUSSION

Our results show that a linear model using low-level sound features can predict the neural responses to naturalistic speech in the perisylvian regions. This result was observed with both ECoG and fMRI. Exploiting the fine-grained temporal resolution of ECoG, we also found a gradient of tuning to the temporal characteristics of speech spreading from posterior STG to IFG. Moreover, IFG was tuned to coarse acoustic features of speech sounds at a temporal rate close to or slower than the rate of word transitions.

### Information propagation in speech perception: ventral pathway from posterior STG to IFG

Low-level feature tuning along the perisylvian cortex was observed in ECoG and reproduced using fMRI. Thus, we confirmed that sparse spatial coverage of ECoG did not bias the encoding results. The present fMRI results were obtained using continuous naturalistic stimuli and resembled the tuning patterns reported previously with more controlled stimuli (Norman-Haignere et al., 2015). In ECoG, the neural responses in posterior STG were modeled better, compared to the more anterior associative cortices, such as IFG. The prediction accuracy dropped gradually, as the electrode location moved anteriorly and away from the posterior STG. Regional differences in the prediction accuracy were associated with tuning to different time lags, with STG being tuned to features at short time lags (< 150 ms) after the sound onset and IFG being tuned to longer time lags (up to 500 ms).

Multiple electrophysiological studies have reported a temporal evolution of the activation from posterior STG to IFG during sentence comprehension (Brennan and Pylkkänen, 2012; Halgren et al., 2002; Kubanek et al., 2013). A recent study reported propagation of acoustic features of speech along the STG (Hullett et al., 2016). The evidence presented here is in line with these findings. However, it also shows that speech-specific low-level auditory features propagate from early sensory areas to distal associative cortex as far as the IFG. Evidence for propagation of low-level sensory features in the brain has previously been reported for both visual and auditory modalities (Bastos et al., 2015; Fontolan et al., 2014). In visual perception, information is transferred from early visual processing areas towards the inferior temporal cortex, where high-level object representations (such as object categories) arise (Güçlü and van Gerven, 2015a; Khaligh-Razavi and Kriegeskorte, 2014). An analogous representational gradient from early processing areas (posterior STG) towards higher-level cortex (IFG) has been reported for auditory perception using neural responses to musical pieces containing lyrics (Güçlü et al., 2016). Our results suggest that a similar form of signal transfer might take place in speech perception.

The present findings are connected to the research on language comprehension along the perisylvian cortex through a dual-stream theory of language processing (Hickok and Poeppel, 2004; Rauschecker, 1998). In particular, these findings are in line with the view that there is a ventral pathway of language processing in the brain (Hickok and Poeppel, 2004; Rauschecker and Tian, 2000; Saur et al., 2008). Along this pathway the information is passed from posterior STG to anterior STG and IFG. The ventral stream is thought to accommodate transition of sound to meaning.

### Encoding of low-level features of speech along the ventral pathway: temporal encoding gradient

It has so far been challenging to uncover how exactly the sound-to-meaning mapping is implemented in the brain. Here we investigated which specific features in the auditory input were most relevant for modeling the neural responses in different brain areas along the perisylvian cortex. In an effort to interpret the tuning preferences in different cortical regions we explored how low-level TMs mapped onto the temporal rates of various language units.

Multiple studies have previously shown a relationship between the neural responses in posterior STG and the spectrotemporal features of sound (Hullett et al., 2016; Martin et al., 2014; Pasley et al., 2012; Santoro et al., 2014; Schönwiesner and Zatorre, 2009). Here, with both ECoG and fMRI, we found that the cortical sites along the pathway from posterior STG to IFG showed tuning to coarse SMs (<1 cyc/oct). Coarse SMs represent a spread of energy over a broad range of frequencies (Chi et al., 2005). Such modulations are characteristic of formants and their transitions (Elliott and Theunissen, 2009). Using behavioral experiments (Drullman, 1995; Elliott and Theunissen, 2009; Fu and Shannon, 2000), it was shown that coarse SMs (<1 cyc/kHz) are crucial for speech intelligibility. In our audio data speech also exhibited more energy along the dimension of TMs, compared to SMs, where the energy concentrated in the range of 0.25 – 4 cyc/oct (**Figure 2.1b**). Notably, in the case of modeling neural responses to music, we observed tuning to both fine and coarse SMs.

TMs capture the dynamics in the audio signal. The energy along a specific frequency range can decline rapidly over time, and thus occur at a fast temporal rate; or last for longer periods of time, and thus occur at a slower temporal rate (Chi et al., 2005). The encoding of the temporal organization of speech in the ECoG responses showed distinct profiles along the ventral pathway. The TM tuning profiles in STG were in accordance with the previous findings (Hullett et al., 2016). At the same time, the present results extend more posteriorly (posterior STG: TMs > 10 Hz), and more anteriorly, towards IFG. We show that IFG is tuned to even slower TMs, compared to anterior STG (0.5 – 4 Hz). Based on perceptual experiments, a number of studies have previously linked speech intelligibility to TMs in the range of 0.25 – 30 Hz (Arai et al., 1999; Drullman et al., 1994;

Elliott and Theunissen, 2009). Specifically, various subphonemic features, such as formant transitions and voice onset time, have been linked to TMs in the 30 – 50 Hz range (Giraud and Poeppel, 2012), phoneme identification has been linked to TMs of > 16 Hz (Drullman et al., 1994), and syllable identification has been linked to TMs at 4 Hz (Arai et al., 1999; Houtgast and Steeneken, 1985). Neural tracking of sentence- (1 Hz) and syllable-specific (4 Hz) TMs have been shown using highly controlled synthesized speech stimuli (Ding and Simon, 2012). Here we show that a correspondence between the stimulus TMs and neuronal response preferences can also be observed during naturalistic speech comprehension. Tuning to the TMs of voice onset time, formant transitions, and phoneme, syllable and word boundaries varied in an orderly fashion along the pathway from posterior STG to IFG.

### Role of IFG in continuous speech comprehension

The present results conform to the previously denoted roles of STG and IFG in language processing. Posterior STG is tuned to fast TMs (> 10 Hz) and short time lags (< 150 ms), which allows efficient processing of fine changes in the auditory input, and accommodates phonological processing and phoneme identification (Dehaene-Lambertz et al., 2000; Formisano et al., 2008; Chang et al., 2010). As the information about the input signal passes through the ventral stream of language processing, it becomes represented at coarser temporal rates and the sound representation exhibits longer time lags. Eventually, IFG processes the incoming speech signal at a 300 – 500 ms delay compared to the sound onset. Here, the signal is spectrally spread over more than one octave and temporally over 250 – 1000 ms. A comparison of the TM tuning of IFG to the temporal rate of individual word transitions (**Figure 2.5b**), suggests that this region may be capturing speech-related information integrated over several words.

A subregion of IFG that exhibited tuning to coarse features of speech was also engaged in a different language task (verb generation, same patients). Given that the same neuronal populations appear to be involved in both maintaining coarse representations of the incoming speech and a verb generation task, it becomes more challenging to agree on one specific function of IFG. One influential theory, which attempts to integrate over multiple findings regarding the function of IFG, is the Memory, Unification and Control theory (Hagoort, 2013, 2005). According to this theory, IFG accommodates the general process of unification, or the combining of the building blocks of language information into larger structures (Hagoort, 2013). With respect to the present results, during speech perception IFG appears to be involved in the integration of incoming speech information over windows of 500 ms and the maintenance of a more 'abstract', integrated representation of the input that arises in neuronal populations of IFG 300 – 500 ms after the sound onset.

## Limitations and future directions

The present work has a number of limitations. For instance, the results obtained using ECoG data should be interpreted with care, because ECoG is an invasive method, which only registers neural responses from the cortical surface and not from deep and folded brain regions.

We studied the neural responses to speech using fragments of a feature film. This experimental setup allowed us to obtain neural responses to speech in naturalistic circumstances, without introducing experimental artifacts, and explore the data using bottom-up approaches. In case of fMRI we recovered the general contour of the low-level feature tuning shown in related work (Norman-Haignere et al., 2015; Santoro et al., 2014; Schönwiesner and Zatorre, 2009). However, using continuous naturalistic stimuli and a task-free experimental paradigm did not allow us to uncover the fine-grained tuning profiles in STG reported in these studies. Additionally, the adopted paradigm made it unclear to what extent the observed pattern of neural responses was explained by attention shifts, multisensory integration and other top-down processes, e.g. the McGurk effect, which represents interaction between visual and auditory speech modalities (McGurk and MacDonald, 1976). It is conceivable that our use of a dubbed soundtrack affected the latency of the ECoG responses to speech. However, we believe it is unlikely to have influenced the specific low-level tuning of different cortical regions or their lag difference relative to each other, since the observed tuning profiles are consistent with previous reports on auditory processing (Hullett et al., 2016).

One of the possible future directions of the present work could be to investigate the feedback connections during continuous speech comprehension. Specifically, it is interesting to investigate how our results relate to the predictive coding account of speech perception (Arnal and Giraud, 2012; Hickok, 2012). To address this issue, we would investigate whether the neural responses contain an indication about the features of upcoming speech and take into account oscillatory patterns in lower frequency bands.

## Conclusions

In the present study we investigated neural tuning to low-level acoustic features of speech sounds while attending to a short movie. The present findings support the notion of low-level sound information propagating along the speech perception pathway. In particular, increasingly coarse temporal characteristics of speech are encoded along the pathway from sensory (STG) areas towards anterior associative cortex (IFG). Finally, we observed that the sound features reflecting markers of words and syllable transitions travel as far as the IFG, which is associated with language processing.

## 2.4 MATERIALS AND METHODS

A series of 30-second video fragments from a feature film (Pippi Longstocking, 1969) were concatenated into a coherent storyline. Fragments containing conversations between people were alternated with fragments containing music (no speech) while the visual story proceeded. The resulting movie lasted 6.5 minutes. Speech was dubbed since the original language was Swedish. There were two repeating music fragments; therefore one of them was taken out from all the analyses. Thus, six music and six speech 30-second fragments were analyzed.

The NSL toolbox (Chi et al., 2005) was used to extract low-level spectrotemporal features from the audio signal. The computational model by Chi et al., 2005 is biologically inspired and implements two stages of sound processing: the cochlea stage, during which the audio spectrogram is computed at logarithmically spaced central frequencies in the range 180 – 7000 Hz; and the early cortical stage, during which a non-linear spectrotemporal representation of sound is obtained by filtering the spectrogram with various modulation filters.

The soundtrack of the short movie was downsampled to 16 kHz and a sound spectrogram was obtained at 8 ms frames for 128 logarithmically spaced frequency bins in the range of 180 – 7000 Hz (**Figure 2.1a**). Next, a non-linear modulation based feature representation of the sound was obtained by filtering the spectrogram at every audio frequency bin and time point with a bank of 2D Gabor filters varying in spatial frequency and orientation. The output was a 4D set of low-level audio features: spectral modulations (SMs) × temporal modulations (TMs) × frequency bin × time point. The non-linear features were obtained along the dimensions of SMs over the range of 0.2 –16 cycles per octave (cyc/oct) in linear steps of 0.2 cyc/oct, and TMs over the range of 0.2 –16 Hz in linear steps of 0.2 Hz. The resulting feature set was of size 81 × 162 × 128 × 3750 per each 30-second fragment. The features along the dimension of TMs accounted for upward and downward sweeps in the spectrogram and therefore contained twice as many features.

To quantify the difference in the spectrotemporal features between speech and music fragments, within each 30-second fragment the feature set was averaged along the frequency bins. A 1-second long vector per each combination of SM and TM features was obtained for speech and music, separately, by averaging the SM-TM features over all seconds of all sound-specific fragments. For each combination of SM and TM features, the music and speech 1-second long vectors were compared with a two-tailed t-test. The p-values were set at .001, Bonferroni corrected for the number of SM-TM features (81 × 162, in total).

For the visualization (**Figure 2.1b and c**), within each fragment the original size feature set (81 × 162 × 128 × 3750) was averaged along the frequency bins and time points. Then the SM-TM features were averaged across sound-specific fragments to obtain a single averaged SM-TM feature set for speech and music, separately.

## Experimental design and statistical analysis

### ECoG experiment

### *ECoG setup and movie watching experiment*

Fifteen patients (age 28 ± 12, nine females) with medication-resistant epilepsy underwent subdural electrode implantation to determine the source of seizures and test the possibility of surgical removal of the corresponding brain tissue. All patients gave written informed consent to participate in accompanying ECoG and fMRI recordings and gave permission to use their data for scientific research. The study was approved by the Medical Ethical Committee of the Utrecht University Medical Center in accordance with the Declaration of Helsinki (2013).

All patients were implanted with clinical electrode grids (2.3 mm exposed diameter, inter-electrode distance 10 mm, between 64 and 120 contact points), one patient had a high-density grid (1.3 mm exposed diameter, inter-electrode distance 3 mm). Thirteen patients were implanted with left hemispheric grids. Most had left hemisphere as language dominant (based on fMRI or Wada test). All patients had perisylvian grid coverage and most had electrodes in frontal and motor cortices. The total brain coverage is shown in **Figure 2.8**. Patient-specific information about the grid hemisphere, number of electrodes and cortices covered is summarized in **Table 1**.

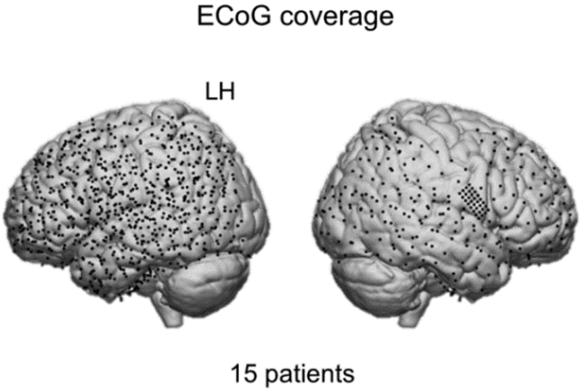


**Figure 2.8. ECoG coverage.** Per patient, the electrode locations were normalized using individual affine transformation matrices obtained with SPM8. The normalized electrode locations were pooled across patients and displayed on the standard MNI brain to show overall brain coverage.

**Table 2.1. Electrode grid information**. Information about number of electrodes, grid hemisphere, covered cortices, handedness and language dominant hemisphere per patient. LH: left hemisphere, RH: right hemisphere, F: frontal, M: motor, T: temporal, P: parietal, O: occipital cortices, R: right, L: left.

| Patient | N of electrodes | Grid hemisphere | Cortices covered | Handedness | Language dominance |
|---|---|---|---|---|---|
| 1 | 96 | LH | F, T, P | R | L (Wada) |
| 2 | 112 | LH | F, M, T, P, O | R | L (fMRI) |
| 3 | 96 | LH | F, M, T, P, O | R | L (Wada) |
| 4 | 104 | LH | F, M, T, P | R | L (Wada) |
| 5 | 48 | LH | F, M, T, P | L | L (fMRI) |
| 6 | 120 | LH | F, M, T | R | L (Wada) |
| 7 | 112 | LH | F, M, T, P | R | L (fMRI) |
| 8 | 64 | LH | F, M, T | R | L (fMRI) |
| 9 | 112 | LH | M, T, P, O | R | L (fMRI) |
| 10 | 64 | LH | M, T, P | R | L (Wada) |
| 11 | 96 | RH | M, T, P, O | L | R (Wada) |
| 12 | 88 | LH | T, P, O | R | L (fMRI) |
| 13 | 112 | LH | F, T, P | R | R (Wada) |
| 14 | 120 | RH | F, M, T | R | L (Wada) |
| 15 | 96 | RH | F, T, P, O | L | L (Wada) |

In the movie watching experiment each patient was asked to attend to the short movie made of fragments from Pippi Longstocking (1969). The video was delivered on a computer screen (21 inch in diagonal). The stereo sound was delivered through speakers whose volume level was adjusted for each patient.

During the experiment ECoG data were acquired with a 128 channel recording system (Micromed, Treviso, Italy) at a sampling rate of 512 Hz filtered at 0.15–134.4 Hz. The movie was presented using Presentation® (Version 18.0, Neurobehavioral Systems Inc) and sound was synchronized with the ECoG recordings.

### *ECoG data processing*

All electrodes with noisy or flat signal (visual inspection) were excluded from further analyses. After applying a notch filter for line noise (50 and 100 Hz) common average re-referencing was applied separately for clinical and high-density grids. Data were transformed to the frequency domain using Gabor wavelet decomposition at 1 – 120 Hz in 1 Hz bins with decreasing window length (4 wavelength full-width at half max, FHWH).

Finally, high frequency band (HFB) amplitude was obtained by averaging amplitudes for the 60 – 120 Hz bins, and the resulting time series per electrode were downsampled to 125 Hz.

Electrode locations were coregistered to the anatomical magnetic resonance image in native space using computer tomography scans (Hermes et al., 2010) and Freesurfer (http://surfer.nmr.mgh.harvard.edu/). The Desikan-Killiany atlas (Desikan et al., 2006) was used for anatomical labeling of electrodes (closest cortical structure in the radius of 5 mm). All electrode positions were projected to Montreal neurological institute (MNI) space using SPM8 (Welcome Trust Centre for Neuroimaging, University College London, UK).

## Neural tuning to low-level sound features in ECoG

### *Feature space*

To model ECoG HFB responses, features describing SMs, TMs, and time lag were used. To reduce the computational load, reduced spectrotemporal feature sets were obtained. The reduced feature set included 17 features spaced logarithmically along the spectral dimension (0.03 to 8 cyc/oct) and 17 features spaced logarithmically along the temporal dimension (0.25 to 64 Hz). The resulting feature set was of size 17 × 34 × 128 × 3750 per sound-specific fragment. The features along the dimension of TMs accounted for upward and downward sweeps in the spectrogram and therefore contained twice as many features. The features were averaged along 128 frequency bins, as well as along upward and downward sweeps, resulting in a 17 × 17 × 3750 feature set per fragment. Due to a possible delay of the neural responses with respect to the audio onset, all negative time lags up to -500 ms relative to the audio onset were added to the feature set as the third feature dimension. Thus, to model a neural response at time t, all SMs and TMs within [t-500, t] were used. At the sampling rate of 125 Hz, this resulted in 63 time lags. The resulting feature matrix had 17 × 17 × 63 features per time point.

### *Encoding model*

Kernel ridge regression (Murphy, 2012) was used to model responses of an electrode ($\mathbf{y}_e$) as a linear combination of nonlinear stimulus features ($\mathbf{X}$):

$$\mathbf{y}_e = \boldsymbol{\beta}_e{}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\varepsilon}_e \tag{2.1}$$

where $\boldsymbol{\varepsilon}_e \sim \mathcal{N}(0, \sigma^2)$.

An $L^2$ penalized least squares loss function was analytically minimized to estimate the regression coefficients ($\boldsymbol{\beta}_e$), and the kernel trick was used to avoid large matrix inversions in the feature space:

$$\boldsymbol{\beta}_e = \mathbf{X}^{\mathrm{T}}(\mathbf{K} + \lambda_e \mathbf{I}_n)^{-1}\mathbf{y}_i \tag{2.2}$$

where $\mathbf{X}$ and $\mathbf{y}_e$ is an estimation set with $n$ data-cases and matrix $\mathbf{K}$ is the Gram matrix:

$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}} \tag{2.3}$$

A nested cross-validation was used to estimate the complexity parameter that controls the amount of regularization ($\lambda_e$) (Güçlü and van Gerven, 2014). First, a grid of the effective degrees of freedom of the model fit was specified. Then, Newton's method was used to solve the effective degrees of freedom for $\lambda_e$. Finally, $\lambda_e$ that resulted in the lowest nested cross-validation error was taken as the final estimate.

### *Model evaluation*

A five-fold cross-validation was used to validate the model performance. The validation procedure was used to tackle overfitting, i.e. to obtain results generalizable to novel audio data. The model was trained using data from music and speech fragments combined. We tested the model on predicting neural responses to music and speech separately. Each of the sound-specific 30-second fragments was partitioned into five 6-second chunks. When testing the model on speech data, one chunk was selected per speech fragment to make up a test set. In each cross-validation fold different chunks were selected per fragment, so that no data points were shared in test datasets across five folds. The first second of each chunk used as a test set was discarded because the preceding lag information during the chunk's first 500 ms was not part of the test set. The five truncated chunks were concatenated and constituted the test set. The remaining data from both speech and music fragments were used for training. The data immediately following the chunk used in the test set (up to one second) were also excluded from the training set, as the lag information for these data were used in the test set. Thus, no data points were shared between train and test datasets in one cross-validation fold. The procedure was identical when selecting test data from music fragments to test the model on music data.

Model performance was measured as the Spearman correlation between predicted and observed neural responses in the held-out test set. The correlation values were averaged across five cross-validation folds and were transformed to t-values for determining significance (Kendall and Stuart, 1973). The correlation values reported here were significant at $p < .001$, Bonferroni corrected for the number of electrodes. Throughout the manuscript, we report the p-value threshold prior to its correction for multiple comparisons. Thus, the actual p-value threshold is smaller than the reported one. In case of the Bonferroni correction, the actual p-value threshold is the reported threshold over the number of comparisons.

### *Retraining the encoding model with varying amounts of training data*

We retrained the encoding model using smaller training data sets: 10%, 20%, 40%, 60% and 80% of all available training data set to investigate the sufficiency of the amount of training data in the encoding analysis. The model was retrained to predict the neural responses to speech. The prediction accuracy obtained using different amounts of training data was compared. The results for all 1283 electrodes are shown in **Figure 2.9a**. Next, the 130 electrodes with significant model performance in speech were selected. Their prediction accuracy obtained using different amounts of training data was compared using one-way ANOVA test (F = 65.57, p = 4.1 × 10-57). Tukey's honest significance difference criterion was used to correct for multiple comparisons. Mean prediction accuracy was significantly lower when using 10% training data, compared to all other models (p = 2.07 × 10-8 for all comparisons). Mean prediction accuracy for models using 60%, 80% and 100% of training data did not differ significantly: p = .63 (60% vs 80%), p = .21 (60% vs 80%) and p = 0.98 (80% vs 100%). The details are summarized in **Figure 2.9b**. The results of the model comparison suggested leveling out of the mean prediction accuracy over electrodes tuned to speech starting at 60% of training data towards using 100% of training data. Altogether, this indicates that the amount of training data in this study was sufficient for robust assessment of the speech encoding model.

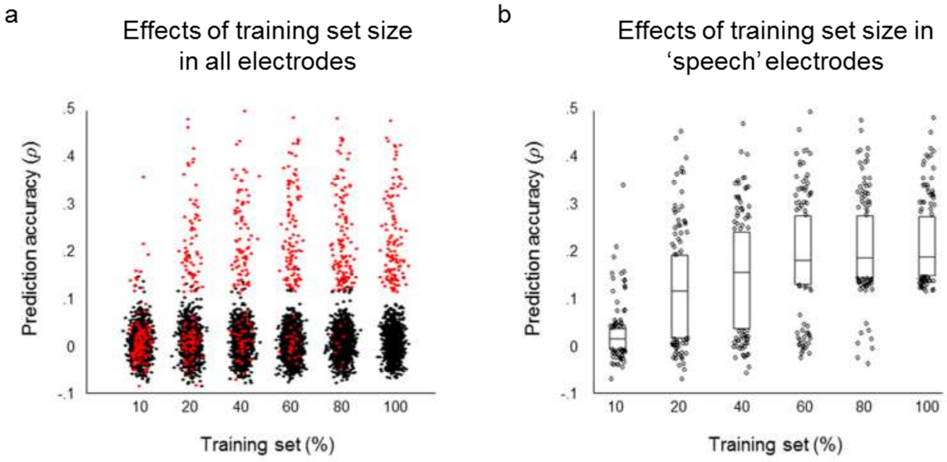


**Figure 2.9. Effects of training size on model performance when tested on speech in ECoG. (a)** Dot plot of prediction accuracy in all electrodes (N = 1283). Red dots indicate electrodes with significant model performance when trained on 100% training data and tested on speech (130 electrodes). Black plots are remaining electrodes. **(b)** Box plots of prediction accuracy in 130 electrodes with significant model performance when trained on 100%

training data and tested on speech. The lower and upper boundaries of the box plots show 25th and 75th quantile, respectively. The horizontal bars show the median. The dots are the remaining accuracy values. One-way ANOVA results showed significant difference between results obtained with 10% training data and all other groups. No significant difference in prediction accuracy was revealed between results obtained with 60%, 80%, or 100% of all training data.

### *Anterior-posterior gradient of model prediction accuracy*

A one-way ANOVA test was used to determine the effect of the anatomical label of the electrode on the model prediction accuracy in speech. Five cortical regions along the IFG – STG axis were chosen: IFG pars triangularis and pars opercularis, precentral gyrus, postcentral gyrus and STG. Temporal pole and pars orbitalis of IFG were not included because each only comprised one electrode with significant model performance. Each of the five chosen cortical regions comprised five or more electrodes. Across patients, 102 electrodes were included in the analysis. Given that the F statistic was significant at p = .02, the least significant difference test was applied to establish which cortical regions had significantly different mean prediction accuracy.

Additionally, a linear regression was fit to model the prediction accuracy in speech per electrode ($r_e$) based on the electrode's anatomical label (eq. 4). The five previously considered anatomical labels were replaced by integer values (from 1 to 5; $l_e$) based on the distance of the label from STG, thus for the electrodes in STG, $l_e = 1$, whereas for the electrodes in pars opercularis, $l_e = 5$:

$$r_e = \boldsymbol{\beta}_e{}^{\mathbf{T}}\mathbf{l}_e + \varepsilon_e \tag{2.4}$$

where $\mathbf{l}_e$ is a vector $[1, l_e]^{\mathbf{T}}$.

An F-statistic comparing the model fit using the five anatomical labels against the constant model was computed. The estimated regression coefficients for the label predictor and the intercept are reported. An F-statistic comparing the model fit using the label predictor against the constant model was computed.

## Distinct tuning profiles across cortical sites in ECoG

### *Clustering of regression coefficients*

The learned regression coefficients were selected for the electrodes with significant model performance (130 electrodes for speech and 47 electrodes for music). Per electrode, the regression coefficients were z-scored over the features. To improve display of results the regression coefficients were also z-scored over the selected electrodes, however comparable clustering results were obtained without z-scoring over the electrodes.

Affinity propagation clustering (Frey and Dueck, 2007) was applied to the regression coefficients across the electrodes to obtain clusters of electrodes with similar feature tuning profiles. The 3D regression coefficient matrices (SMs × TMs × time lags) were vectorized. Euclidian distance was used as a measure of similarity between resulting regression coefficient vectors. The grouping criterion, or similarity preference parameter ($p$), was set to one of the default estimates (Frey and Dueck, 2007):

$$p = \frac{5\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} D_{ij}}{m}\right)^2 + \min_{i,j} D_{ij}}{6} \tag{2.5}$$

where **D** is the Euclidian distance matrix for the regression coefficient vectors and $m$ is the number of electrodes. These parameters produced better visualizations, however various parameter setting were also tried.

Affinity propagation clustering determined the optimal number of clusters automatically, based on the similarity preference parameter ($p$).

For the electrodes with significant model performance in speech, the analysis yielded six clusters which comprised electrodes from about a half of all patients. Altogether, these clusters contained 81% (105 electrodes) of all electrodes, used in the clustering analysis (130 electrodes). Of the remaining clusters, seven clusters only comprised electrodes from one patient, and two clusters comprised electrodes from three patients (altogether, 25 electrodes). For the electrodes with significant model performance in music, the analysis yielded three clusters comprising 100% of all electrodes used in the clustering analysis (47 electrodes).

For subsequent analyses the three feature dimensions (SM, TM, time lag) were separated and averaged across electrodes within each cluster. To obtain tuning profiles for each dimension, regression coefficients along the other two dimensions were averaged. Significance of the resulting tuning profiles was assessed by permutation testing. The procedure was performed on the regression coefficients, averaged over all cluster members. The electrode-to-cluster assignments were permuted 10000 times, resulting in a null distribution of tuning profiles for each feature along each feature dimension: 17 for SMs, 17 for TMs and 63 for time lags. The original cluster tuning profiles were then evaluated in each feature relative to those null distributions, yielding exact statistical p-values. The cluster mean regression coefficients reported here were significant at p < .01, Bonferroni corrected for the number of features along each feature dimension.

### *Stepwise linear regression to relate tuning profiles to prediction accuracy*

A stepwise linear regression as implemented in MATLAB 2016a (The MathWorks Inc., Natick, MA, United States) was used to model prediction accuracy as a function of time lag, TM and SM tuning. The regression was fit on 130 electrodes with significant model performance in speech. For each electrode maximal regression coefficients for time lag, TMs and SMs were calculated and used as predictors of the electrode's model prediction accuracy ($r_e$). The full-interaction model was specified where apart from the individual predictors (lag, TMs, SMs), products of all their combinations were added to the regression:

$$r_e = {\beta_e}^{(0)} + \sum_{i=1}^{3} {\beta_e}^{(i)} {w_e}^{(i)} + \sum_{i=1}^{2} \sum_{\substack{j=2,\\ j\neq i}}^{3} {\beta_e}^{(ij)} {w_e}^{(i)} {w_e}^{(j)} + {\beta_e}^{(123)} \prod_{i=1}^{3} {w_e}^{(i)} + \varepsilon_e \tag{2.6}$$

where ${\beta_e}^{(0)}$ is the bias term, and ${w_e}^{(i)}$ is the electrode's maximal regression coefficient for each of the features $(i)$: time lag, TMs and SMs.

Based on the analysis of the residuals, the algorithm determined which predictors contributed significantly to the model performance. The estimated regression coefficients for the significant predictors are reported. An F-statistic comparing the model fit using significant predictors against the constant model was computed.

### *Retraining the encoding model with reduced lag representations*

To further quantify the effects of time lag on the model performance and electrode-to-cluster assignment (six clusters in speech), the model was retrained using reduced lag representations, and tested to predict neural responses to speech. In the first case, the model was retrained without using any lag information, in the second case, the model was retrained to predict the neural responses to speech only using sound features at a lag of -500 ms. The resulting prediction accuracy produced by varying input lag information (*Lag0-500*, *Lag0* and *Lag500*) were compared using a two-way ANOVA test with unbalanced design (due to varying number of electrodes per cluster). The two main factors in the test were the time lag (0 – 500, 0 or -500) and the cluster assignment (six speech clusters). The interaction between the main factors was also estimated. Given that the F statistic was significant at p < .01, the least significant difference test was applied to establish which models had significantly different mean prediction accuracy per cluster.

Additionally, for each cluster a one-way ANOVA test was performed to quantify the difference in the SM and TM tuning learned in all three cases. For *Lags0-500* the regression coefficients were averaged over time lags. For each test, cluster average tuning profiles were used. Only electrodes with significant model performance in speech (130 electrodes) were considered.

### *Relation to language features*

Annotation of the movie sound track with linguistic markers was done manually using Praat (Boersma and Weenink, 2016, from http://www.praat.org/). Five linguistic features were considered: voice onset time (VOT), formant transitions, as well as phoneme, syllable and word boundaries. For each linguistic feature, we calculated the duration of each instance $(d_i)$ measured in ms and converted it to the rate value $(r_i)$ in Hz. The rate $(r_i)$ was obtained by dividing 1000 over the duration $(d_i)$.

## FMRI experiment

### *FMRI subjects and setup*

Eleven patients watched the same short movie during the presurgical fMRI recordings. The video was delivered on a screen through a scanner mirror, and the audio was delivered through earphones.

Functional images were acquired on a Philips Achieva 3T MRI scanner using 3D-PRESTO (Neggers et al., 2008) (TR/TE = 22.5/33.2, time per volume 608 ms, FA = 10°, 40 slices, FOV = 224 × 256 × 160 mm, voxel size 4 mm). Anatomical T1 images were acquired using TR/TE 8.4/3.2 ms, FA = 8°, 175 = slices, FOV = 228 × 228 × 175, voxel size of 1 × 1 × 2 mm.

### *FMRI data preprocessing*

The functional data were corrected for motion artifacts and coregistered with the anatomical scan using SPM8 (Welcome Trust Centre for Neuroimaging, University College London, UK). Four patients were discarded, as they moved more than 4 mm in at least one of the directions. The data were despiked using ArtRepair (Mazaika et al., 2009) at 10% variance, detrended and high-pass filtered at 0.009 Hz using mrTools (http://gru.stanford.edu/doku.php/mrTools/download). Only voxels in grey matter voxels were analyzed. No smoothing was applied prior to encoding or clustering analyses.

### *FMRI encoding model and clustering of regression coefficients*

The 4D SM-TM feature set used in ECoG was adjusted to accommodate analysis with the lower sampling rate of fMRI (to 4 samples/s). Time lag was not included. Based on the previous findings (Santoro et al., 2014), the reduced frequency dimension was added to the feature set, resulting in 17 × 17 × 4 (SMs × TMs × frequency) features per time point. Each feature was convolved with the classical hemodynamic response function (spm_hrf) using default settings. For each patient, six motion regressors were added to the feature set. The model fitting procedures were identical to the model fit on the ECoG neural responses. The split into training and test set was done in the same way, except that no time lag information was used and therefore no truncation of chunks in sound-specific fragments was applied. No data points were shared between the train and test datasets in one cross-validation fold. No data were shared in the test datasets across the five cross-validation folds. Individual prediction accuracy was thresholded at p < .05, FDR-corrected. For the display purposes, the individual prediction accuracy maps were smoothed at FHWH = 10 mm.

Voxels with similar tuning profiles were clustered using affinity propagation (Frey and Dueck, 2007), separately for speech and music. Various clustering parameters were used. Significance of the cluster-specific tuning profiles was assessed by permutation testing. The voxel-to-cluster assignments were permuted 10000 times, resulting in a null distribution of tuning profiles for each feature along each feature dimension. The original cluster tuning profiles were then evaluated in each feature relative to those null distributions, yielding exact statistical p-values. The cluster tuning profiles reported here were significant at p < .01, Bonferroni corrected. The brain maps with clustering results were created only for the purpose of visualization and no explicit group-based statistical analysis was performed. The histogram in **Figure 2.7B** shows that the cluster map was not biased by one patient and included voxels from almost all patients. In order to project the individual results on the standard MNI template, the individual maps (e.g. maps of clustering of regression coefficients) were normalized using DARTEL toolbox (Ashburner, 2007) and an aggregated mask over patients was created in the MNI space. For display purposes the cluster maps were smoothed at FHWH = 10 mm.

### *Agreement between ECoG and fMRI speech encoding results*

To estimate the voxels, corresponding to each ECoG electrode, a sphere was defined in individual fMRI prediction accuracy maps using a 10 mm radius around each electrode's center coordinate. Voxels within that sphere (only grey matter) were associated with the corresponding electrode. No voxel was associated with more than one electrode. Pearson correlation coefficients were calculated between prediction accuracy in ECoG electrodes and corresponding fMRI voxels. The voxel with maximal correlation was determined per sphere. We used prediction accuracy values, uncorrected for significance, as we assumed that low prediction accuracy in both fMRI voxels and ECoG electrodes were indicative of a consistency of the results. Significance was assessed by permuting the prediction accuracy in ECoG, and therefore disrupting the spatial association between ECoG electrodes and fMRI voxels. The permutations were conducted 10000 times per each electrode-voxel pair, and a null distribution of the Pearson correlation coefficients was obtained. The statistical p-values of the original Pearson correlation coefficients were determined and Bonferroni corrected for the number of electrodes per patient.

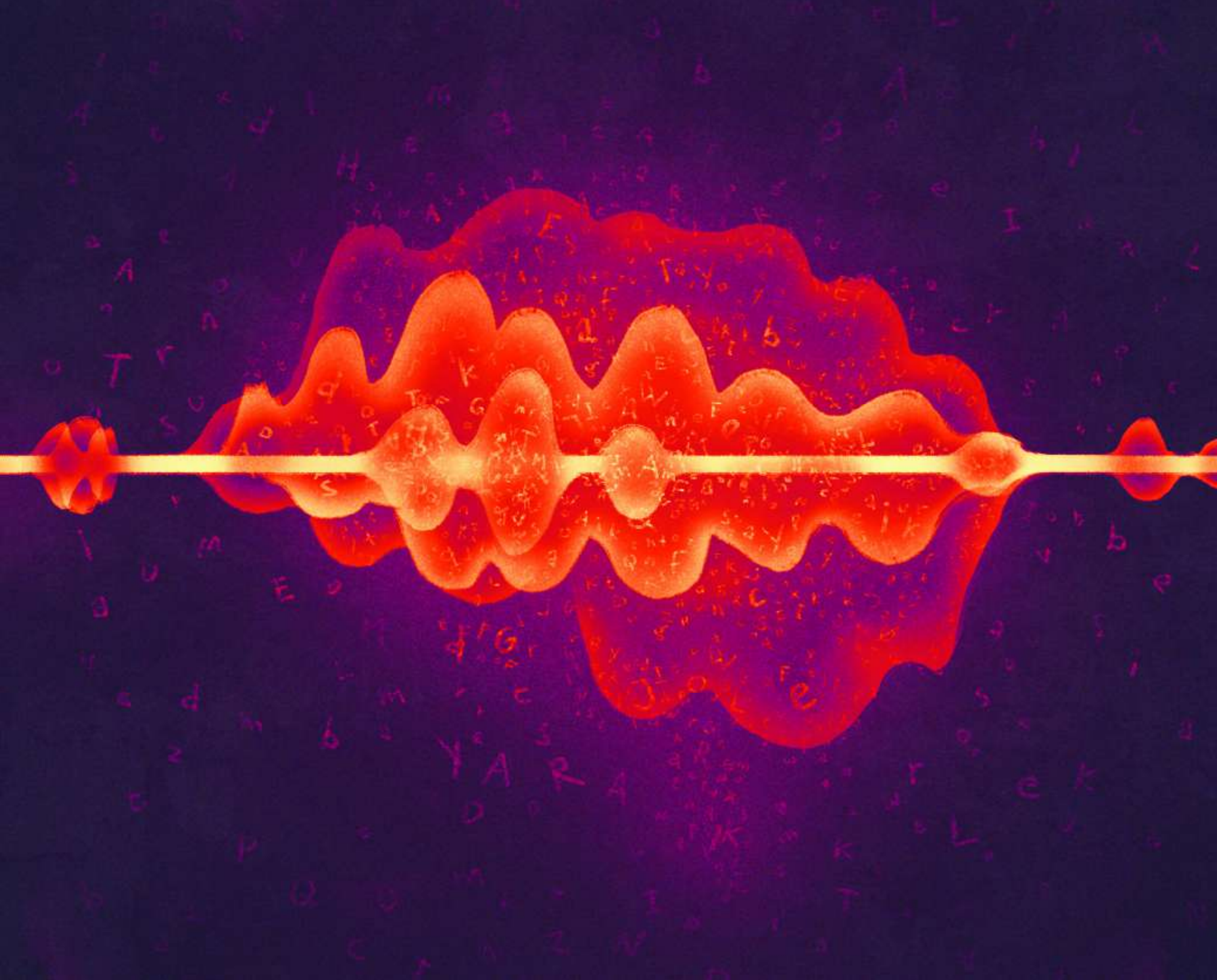
Additionally, the correspondence between ECoG-based and fMRI-based feature tuning profiles, or regression coefficients, was quantified. Per patient, we correlated regression coefficients between electrodes with significant model performance and corresponding fMRI spheres using the Pearson correlation coefficient. The voxel with maximal correlation was determined per sphere. Significance was assessed by permuting the regression coefficients in ECoG 10000 times and obtaining a null distribution of the Pearson correlation coefficients. The statistical p-values of the original Pearson correlation coefficients were determined and Bonferroni corrected for the number of electrodes per patient. Per patient, the Pearson correlation coefficients, significant at p < .001, Bonferroni corrected, were averaged over all electrodes used in the analysis per patient. The averaged Pearson correlation coefficient was weighted by the percentage of electrodes, shown significant in this analysis per patient.

The agreement between ECoG and fMRI encoding results was quantified for both speech and music, however, only speech results showed significant agreement.

chapter

3

# Brain-optimized extraction of complex sound features that drive continuous auditory perception



# CHAPTER 3. BRAIN-OPTIMIZED EXTRACTION OF COMPLEX SOUND FEATURES THAT DRIVE CONTINUOUS AUDITORY PERCEPTION[2]

Understanding how the human brain processes auditory input remains a challenge. Traditionally, a distinction between lower- and higher-level sound features is made, but their definition depends on a specific theoretical framework and might not match the neural representation of sound. Here, we postulate that constructing a data-driven neural model of auditory perception, with a minimum of theoretical assumptions about the relevant sound features, could provide an alternative approach and possibly a better match to the neural responses. We collected electrocorticography recordings from six patients who watched a long-duration feature film. The raw movie soundtrack was used to train an artificial neural network model to predict the associated neural responses. The model achieved high prediction accuracy and generalized well to a second dataset, where new participants watched a different film. The extracted bottom-up features captured acoustic properties that were specific to the type of sound and were associated with various response latency profiles and distinct cortical distributions. Our results also support and extend the current view on speech perception by demonstrating the presence of temporal hierarchies in perisylvian cortex and involvement of cortical sites outside of this region during audiovisual speech perception.

[2] This chapter is based on Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A., & Ramsey, N. F. Brain-optimized extraction of complex sound features that drive continuous auditory perception. *In review.*

## 3.1 INTRODUCTION

Our understanding of how the human brain processes auditory input remains incomplete. In general, our aim is to identify the features that different cortical regions extract from the incoming sound signal, and to understand how they are transformed into high-level representations specific to sound type (speech, music, noise, etc.).

Low-level acoustic sound properties, captured as audio envelope, spectrogram representation, gammatone-filtered representation (Patterson et al., 1987), spectrotemporal modulations (Chi et al., 2005) and related measures, have proven quite effective in supporting development of neural encoding models of both synthetic and natural sound perception (Berezutskaya et al., 2017b; Giraud et al., 2000; Martin et al., 2017; Overath et al., 2008; Santoro et al., 2014). Addressing higher-level features has also been attempted in neural encoding models of sound processing (Mesgarani et al., 2014; Nakai et al., 2018), but higher levels of auditory processing are generally more difficult to model because their characteristics (e.g. in speech or music) remain a topic of theoretical investigation. Higher-level features typically require some form of interpretation and labelling that is based on theoretical constructs and therefore may not match cortical representations. Moreover, little is known about the mechanisms underlying the transition from lower- to higher-level auditory processing, leaving these levels of explanation disconnected.

A possible alternative to using predetermined stimulus features to investigate sound processing is construction of a hierarchical model of auditory processing that can learn the optimal multi-level auditory features from the data (bottom-up), with a minimal set of theoretical assumptions about the higher-level features. Recent advances in deep learning have shown a possibility of construction of such models in the field of object recognition, text processing and sound generation. Recent work in computational neuroscience has indicated that the multi-level features captured in artificial neural networks relate to the hierarchy of cortical representations in visual perception (Güçlü and van Gerven, 2015a; Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016), music (Güçlü et al., 2016) and language (Jain and Huth, 2018) processing.

There has been a long line of research showing evidence of the hierarchical organization of information expressed in both human behavior (Botvinick, 2008; Greenfield, 1991) and the principles of neural processing in perception (Baldassano et al., 2017; DiCarlo et al., 2012; Honey et al., 2012; Moutoussis and Zeki, 1997), action (Grafton and de C. Hamilton, 2007; Loeb et al., 1999) and memory (Hasson et al., 2015). It is possible that the hierarchical structures in the neural responses themselves can be used to directly drive the construction of the encoding model of auditory processing.

Thus, in this study, we sought to develop an encoding model of auditory processing while injecting as little top-down theoretical assumptions or constraints as possible. We modeled the neural responses to sound directly by training a deep artificial neural network on the raw audio input. We assumed that the mechanism of auditory input processing generalizes across individuals, and that models derived from one experimental setting should perform equally well in another. We thus obtained a model from one dataset and applied it to data from new participants in a different set of stimuli.

To this end, we collected neural responses to a full-length feature film from six patients implanted with intracranial electrodes for diagnostic purposes. A deep artificial neural network was trained on the raw audio soundtrack to predict the associated brain responses in all covered regions directly through extraction of multi-layer representations of sound. The obtained model was then applied to data from new participants watching a different movie. Our results indicate a hierarchical processing structure in the brain during auditory perception that is generic across individuals and experimental material (different films and different participants). Moreover, the sound features extracted by the data-driven model proved to capture acoustic properties specific to different types of sound. Crucially, without relying on any theoretical constraints the model learned to incorporate one of the most important characteristics of speech – its hierarchical temporal organization.

## 3.2 RESULTS

The goal of the present study was to construct an auditory perception neural model by predicting the cortical brain responses to the auditory input directly and in a bottom-up manner, injecting as little theoretical constraints to our encoding model as possible. For this, we collected electrocorticography (ECoG) data from six patients who watched a long-duration feature film (Movie I, 78 minutes). Next, a deep artificial neural network was trained on the raw soundtrack of the movie to predict the associated ECoG responses in the high frequency band (HFB, 60-95 Hz)(Crone et al., 1998). We then confirmed that our brain-optimized neural network (BO-NN) model could be successfully applied to a dataset of different participants watching a different audiovisual film (Movie II). The BO-NN model extracted features that better fitted the neural responses in the perisylvian cortex compared to commonly used theory-driven features (spectrotemporal modulation features). Finally, identification and visualization of the key features of the BO-NN model showed presence of features specific to certain sound types, which were also associated with distinct cortical locations, acoustic sound properties and latencies of response to sound onset.

### BO-NN model performance (6 participants, Movie I)

The BO-NN model was designed with two separate streams of convolutional layers (temporal and spectral representations of audio input), the outputs of which were combined and passed to a recurrent module. This model architecture was optimal based on additional tests (see Methods for details). The Spearman correlation between predicted and observed HFB responses was the used performance metric (**Figure 3.1a**). For each out of ten cross-validation folds, the correlation was calculated for the entire left-out test dataset of Movie I (10% of the movie: ~7.8 min, 23,620 time points) and averaged over all folds. Performance did not differ between types of sound (speech, music, noise and ambient sounds) when analyzing them separately (**Figure 3.1b**). Highest accuracy was achieved for electrodes in superior temporal gyrus (STG), reaching $\rho_{max} \approx .5$, compared to inferior frontal gyrus (IFG), precentral, postcentral and middle temporal gyrus (MTG) areas ($\rho_{max} \approx .3$, **Figure 3.1c**). In subjects with substantial dorsal lateral occipitotemporal complex (LOTC) coverage (e.g. S1, S5, S6) many electrodes were also predicted significantly, indicating presence of audiovisual interactions in naturalistic stimuli. Note that there was no coverage in the transverse temporal gyrus.

### BO-NN model performance with unseen data (29 new participants, Movie II)

Performance of the BO-NN model was further validated on a separate movie watching dataset from 29 new ECoG patients (Movie II, 6.5 minutes with 13 interleaved 30-second

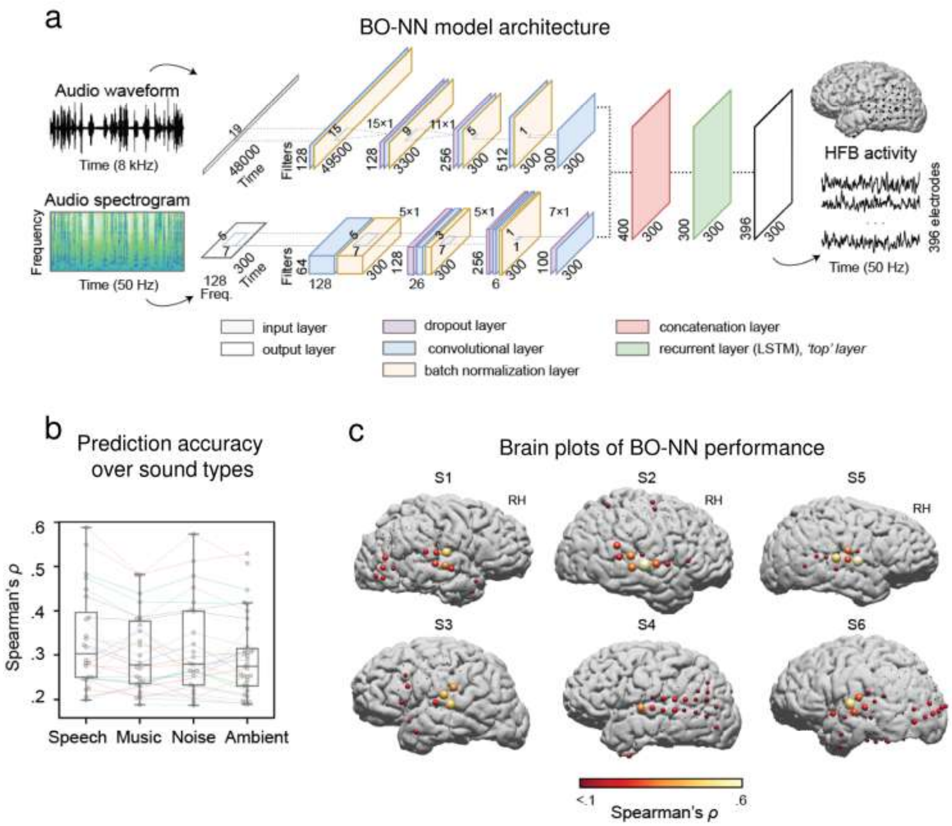


**Figure 3.1. BO-NN model performance with Movie I (six participants). (a)** Architecture of the BO-NN model. The model received two inputs: 1D time-domain audio signal at 8 kHz (audio waveform) and a 2D time-frequency signal at 50 Hz (audio spectrogram). Six-second audio chunks were input at once (48,000 and 300 time points, respectively). Two convolutional streams were trained to process each input separately. Then both representations were concatenated and passed to a recurrent layer. The output was a prediction of the HFB neural signal at 50 Hz (300 time points over 396 electrodes). **(b)** Cross-validated BO-NN model performance (over ten folds) estimated as a Spearman correlation ($\rho$) between the predicted and observed HFB responses on test set of Movie I. Spearman's $\rho$ is shown per type of sound with a sufficient amount of data (at least 10% of the soundtrack): speech, noise, music and ambient sounds. In each test fold all fragments of the same sound type were concatenated, then a Spearman's $\rho$ was calculated per each sound type and averaged over ten cross-validation folds. Each dot is an ECoG electrode with a significant fit at $p < 1 \times 10^{-10}$, Bonferroni corrected for number of electrodes and folds. Boxes show 25th and 75th percentiles, caps show 5th and 95th percentiles. A solid line within each box shows the median. The colored lines connect the model performance for the same electrodes across the sound types. The colored line are less visible for smaller $\rho$ values, however it is clear from the plot that five electrodes ($\rho \geq .4$) are associated with comparable prediction accuracy across different sound types. **(c)** Projection of electrodes with significant CV performance (Spearman's $\rho$ for ~7.8 min of data) on individual cortical

surfaces. Spearman's $\rho$ significant at $p < 1 \times 10^{-10}$, Bonferroni corrected for number of electrodes and folds is shown.

blocks of speech and music). The soundtrack was passed through the pretrained model with all weights fixed, to obtain BO-NN activations. A regularized linear regression was trained per BO-NN layer to predict ECoG responses. Activations of the top recurrent layer of the model were best at predicting the neural responses compared to other layers (two-sided Wilcoxon signed-rank tests, 14 in total: $\bar{Z} = 14.55 \pm 1.81, p < .001$, Bonferroni corrected for number of layers).

Top BO-NN accuracy on speech fragments in Movie II was comparable to the results obtained in Movie I in terms of prediction accuracy and cortical distribution (**Figure 3.2**). Prediction of music fragments was less accurate. Good performance was obtained mostly in perisylvian cortex for both Movie datasets (note that transverse temporal gyrus was not covered).

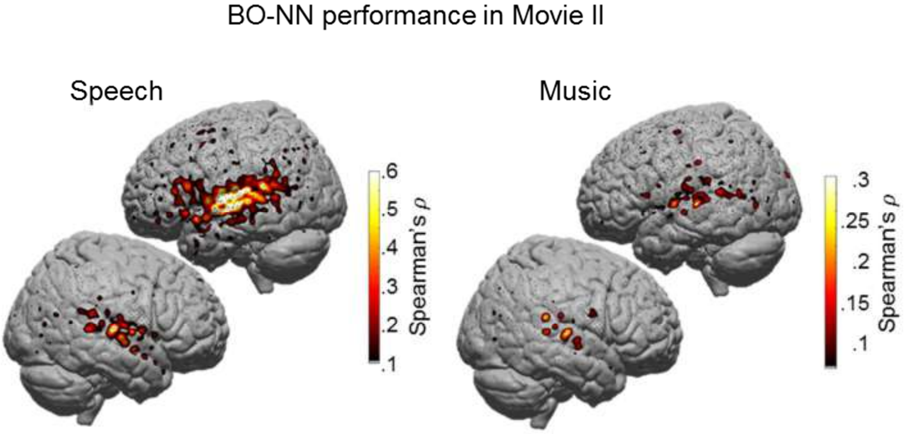


**Figure 3.2. BO-NN model performance with unseen data (29 new participants, Movie II).** CV model performance (Spearman's $\rho$, calculated as with Movie I) is projected to the corresponding electrode locations on the average MNI cortical surface. Individual electrode locations were normalized to the MNI space using patient-specific affine transformation matrices obtained with SPM8. Spearman's $\rho$ significant at $p < .001$ is shown (based on surrogate distribution of shifts). Model performance is shown separately for speech and music test sets. Note the difference in range of Spearman's $\rho$ between speech (up to .6) and music (up to .3). For the visualization purposes, a 2D Gaussian kernel (FWHM = 8 mm) was applied to the coordinate on the brain surface corresponding to the center of the electrode, so that the model performance score (Spearman's $\rho$) faded out from the center of the electrode toward its borders.

Given that the BO-NN model obtained with Movie I data generalized well to Movie II data, and the Movie II dataset contained considerably more electrodes (2218 compared to

396), all further analyses were carried out using data from Movie II. Moreover, using a model that was developed on an independent dataset with separate patients strengthens the generalizability of findings.

### Comparison of the data-driven BO-NN features with control feature sets

Next, we sought to understand how our BO-NN features (top layer) compared to alternative, control features sets using data of Movie II. On the one hand, we sought to compare them against a commonly-used theory-driven feature set, that has been shown to explain considerable variance in the neural responses (Berezutskaya et al., 2017b; Pasley et al., 2012; Santoro et al., 2014). On the other hand, we intended to check that the BO-NN model benefitted from training on the neural responses, and did not offer a good fit simply due to its complex multi-layer architecture. Thus, we compared top-layer BO-NN features to theory-driven spectrotemporal modulation features (STMFs)(Chi et al., 2005), on the one hand, and the top layer of an artificial neural network with the same architecture as the BO-NN model but with randomly initialized weights (Rand-NN). STMFs were extracted from the sound spectrogram using 2D convolutions and were organized along three dimensions: temporal and spectral sound modulations (TMs and SMs, respectively) and frequency. For a fair comparison, we ensured the optimal extraction of STMF features (see Methods for details). Random weights of Rand-NN were sampled from the zero-mean multivariate Gaussian distribution. Given the difference in fitting speech and music responses in Movie II, we carried out all analyses separately for speech and music.

### Sound representation in STG

Data-driven (BO-NN) and theory-driven (SMTFs) feature sets were compared to the neural responses in STG using a representational similarity analysis (RSA, Kriegeskorte et al., 2008a; Nili et al., 2014). RSA allows estimation of how similar the inner structure in one dataset is to the inner structure of other datasets (see Methods for details). Here, for speech, all three feature sets exhibited significant similarity with the representations in STG (one-sided Wilcoxon signed-rank tests: $\bar{Z} = 4.78, p < .001$, Bonferroni corrected for number of feature sets), however top-layer BO-NN representations were most similar to STG (two-sided Wilcoxon signed-rank tests: $Z_{BONN-STMF} = 3.98$ and $Z_{BONN-RANDNN} = 4.78$, both at $p < .001$, Bonferroni corrected for number of feature sets, **Figure 3.3a**). In music however, neither feature set showed better similarity with STG responses. This difference between speech and music conditions also suggests that BO-NN does not simply provide a better fit of the overall shape of the HFB responses (since it was trained to predict neural data whereas STMFs are a theory-driven acoustic model and Rand-NN was untrained). This result likely means that BO-NN captures something about the speech representation STMFs and Rand-NN do not.

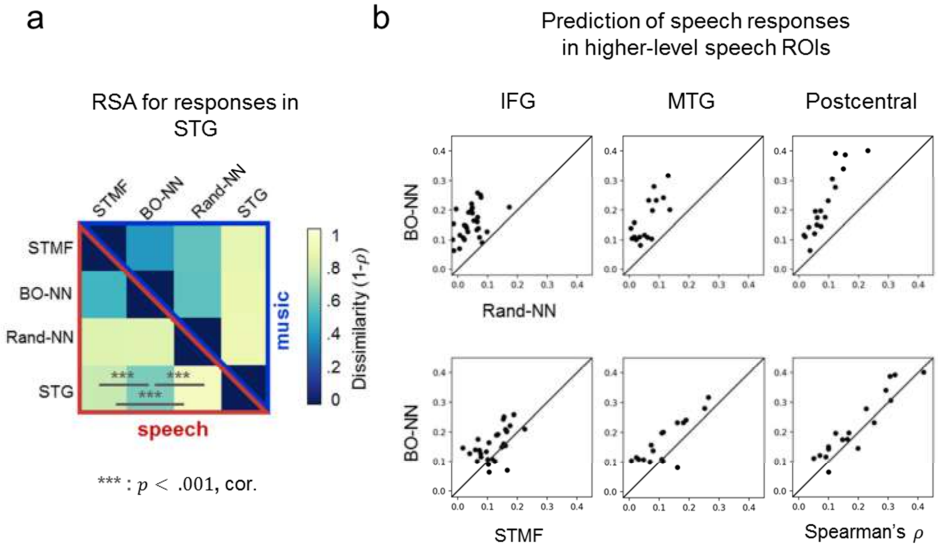


**Figure 3.3. Model comparison.** Comparison of the data-driven BO-NN model trained on the neural responses with two control feature sets on the test set of Movie II. The first control feature set is the theory-driven spectrotemporal modulation features (STMF) model. The second control feature set is the artificial neural network with the same architecture as the data-driven BO-NN but not optimized to fit the neural responses (Rand-NN). **(a)** Second-order dissimilarity matrix showing how dissimilar the audio representations were between STG, data-driven BO-NN model (top layer), theory-driven STMFs model and unoptimized Rand-NN. **(b)** Linear fit per electrode in higher-level speech ROIs: IFG, MTG and postcentral gyrus. Scatter plots in the top panel display the difference in prediction by the data-driven BO-NN model (top-layer) and the theory-driven SMTF model. Scatter plots in the bottom panel display the difference in prediction by the data-driven BO-NN model (top-layer) and the unoptimized Rand-NN model. Each dot is an ECoG electrode with significant Spearman's $\rho$ between predicted and observed HFB responses on test set of Movie II. Spearman's $\rho$ significant at $p < .001$, Bonferroni corrected for number of cortical regions is shown.

## Predictions in higher-level speech regions

Given the observed difference for STG, we further examined responses in higher-level speech processing regions. Three regularized linear regression models were trained to predict the neural responses to speech from BO-NN (top layer,1), STMFs (2) or Rand-NN (3) features. Compared to both STMFs and Rand-NN, top-layer BO-NN reached higher prediction accuracy and fitted more electrodes not only in STG (two-sided Wilcoxon signed-rank test: $Z_{BONN-STMF} = 4.34, p < .001$, +8 sign. els., $Z_{BONN-RANDNN} = 8.94, p < .001$, +49 sign. els.), but also in the higher-level speech processing areas: IFG ($Z_{BONN-STMF} = 3.30, p < .001$, +4 sign. els., $Z_{BONN-RANDNN} = 4.46, p < .001$, +25 sign. els.), MTG ($Z_{BONN-STMF} = 3.01, p = .003$, +5 sign. els., $Z_{BONN-RANDNN} = 3.62, p < .001$, +13 sign.

els.) and postcentral gyrus ($Z_{BONN-STMF} = 2.46, p = .01$, +3 sign. els., $Z_{BONN-RANDNN} = 3.72, p < .001$, +11 sign. els.) (**Figure 3.3b**). This result suggested that the BO-NN model captured features that better explained the neural responses to speech not only in STG but also in higher-level speech processing areas compared to both control feature sets.

## Visualization and interpretation of key BO-NN features

Having observed high prediction accuracy across datasets and subject groups, as well as the improvement over the state-of-the-art theory-driven features in fitting the neural responses to speech, we focused on visualization and interpretation of the BO-NN features. We only explored the top layer of the BO-NN model ("top BO-NN") since these features were best at predicting the ECoG responses. To this end, we used feature visualization, clustering, projection of the associated cortical weight maps ($\beta$-weights of the linear fit of the neural responses from top BO-NN) and relation to the known sound properties to gain insight about each data-driven feature.

### Identification of key BO-NN features

To minimize the number of features to work with, we first identified the key features of top BO-NN using data-driven affinity propagation clustering (Frey and Dueck, 2007). The optimal number of clusters was determined by varying a preference parameter that affects the a priori suitability of each data point to be a cluster center (**Figure 3.4a**, left panel). The optimal estimate of number of clusters (53 key features) falls in the knee of the curve, representing a balance between maximum compression and optimal cluster assignment accuracy (see Methods for details).

Additionally, we verified whether selecting only 53 key BO-NN features, as a result of clustering, still resulted in a high accuracy of predicting the neural responses (**Figure 3.4a**, right panel). Notably, most key features included electrodes from about half of all subjects and thus were not subject-specific.

Next, we evaluated the interpretability of all 53 key BO-NN features using previously mentioned tools (visualization, cortical projections, relation to known sound features, etc.). On that basis we chose a subselection of the most prominent features for more in-depth description and visualization in what follows.

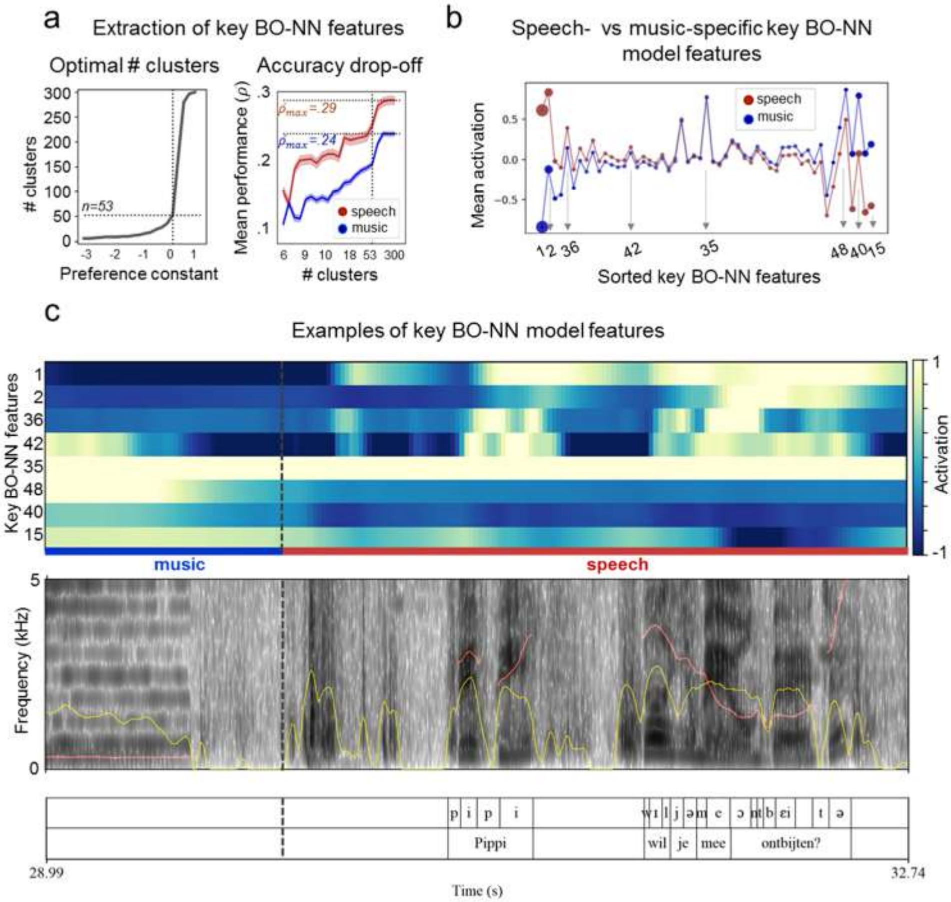


**Figure 3.4. Visualization and interpretation of key BO-NN features. (a)** Optimal choice of a number of key features using AP clustering selected as the knee of the curve over the parameter of *preference* (left panel). Drop-off of prediction accuracy is shown as a function of a number of clusters (right panel). Accuracy for predicting neural responses to speech and music is shown separately. Mean accuracy over all significant electrodes is shown. Shaded areas show the standard error of the mean. **(b)** Key BO-NN features with maximal average activation across speech or music fragments. **(c)** Example ~4 second fragment of activity of selected key features with the corresponding audio spectrogram and language annotations. Top four selected features (#1, 2, 22, 29) were most active during speech blocks, whereas bottom four selected features (#15, 42, 40, 35) were most active during music blocks. Feature activation values are a result of a tanh-transformation and thus are in range of [-1, 1]. Black dotted line showed a border between music and speech blocks. Yellow contour shows sound intensity, red contour shows pitch. Both were extracted automatically from Praat.

## Visualization of key BO-NN features

First, we examined key BO-NN features that showed on average higher activation during either speech or music fragments (**Figure 3.4b**). We visualized the activity of some of these features (**Figure 3.4c**). Overall, multiple key features tracked audio changes visible on the spectrogram, such as silence versus sound moments, changes in intensity and difference between speech and music fragments. Statistical testing showed that some of the key BO-NN features were specific to speech (e.g. key features #1, #2 and #36), some were sensitive to both speech and music (e.g. key feature #42) and some showed enhanced response to music (key features #40, #48 and #15), as tested with two-sided Wilcoxon signed-rank tests: $\bar{Z} = 4.55 \pm .8, p < .001$, Bonferroni corrected for number of key features. One feature exhibited prolonged sustained response during both music and speech fragments, throughout the soundtrack (key feature #35).

## Neural tuning and cortical maps per key BO-NN feature

Next, we examined the cortical weight maps of individual key BO-NN features ($\beta$-weights of the linear fit to the ECoG responses) and observed that multiple features contributed to prediction of the time course of an individual electrode (**Figure 3.5a**).

The cortical map of the highest cortical weights ($\beta$-weights > 2 per each key feature) showed that the perisylvian cortex with adjacent regions were most associated with activity of the key BO-NN features (**Figure 3.5b**). We also visualized the cortical maps for some of the previously mentioned individual features with specific responses to speech and/or music (**Figure 3.5c**). The positive weights expressed a positive contribution of a key BO-NN feature to prediction of the corresponding electrode, whereas the negative weights constitute a reverse relationship.

Electrodes surrounding the transverse temporal gyrus exhibited higher weights for the key feature that was active during both speech and music (key feature #42). Key features #1 and #2 showed distinct maps over the perisylvian cortex and both were most active during the speech fragments. Key feature #15 was more active during the music fragments compared to speech and showed a positive relationship with activity of the electrodes in the dorsal lateral occipitotemporal cortex (LOTC). This result likely reflects the interaction between the visual and auditory streams of Movie II (similar to Movie I). However, this interaction must have occurred at a higher level of conceptual representations given the cortical distribution for this key BO-NN feature and the fact that there was no significant difference in the mean low-level luminance values between the music and speech fragments ($r^2 = .07$, nonsign.), nor was there any difference in the rate of change (derivative) of the luminance values between the two sound types ($r^2 = .02$, nonsign.).

Notably, key feature #2 was associated with negative weights over dorsal LOTC. In fact, the cortical profiles of key feature #2 and key feature #15 seemed to oppose each other. The activation time courses of the two key features were largely anticorrelated during the speech fragments (Spearman's $\rho = -.82$, $p < .001$). During the music fragments the anticorrelation was reduced (Spearman's $\rho = -.41$, $p < .001$), suggesting that the relationship between the key features was modulated by the difference in speech and music.

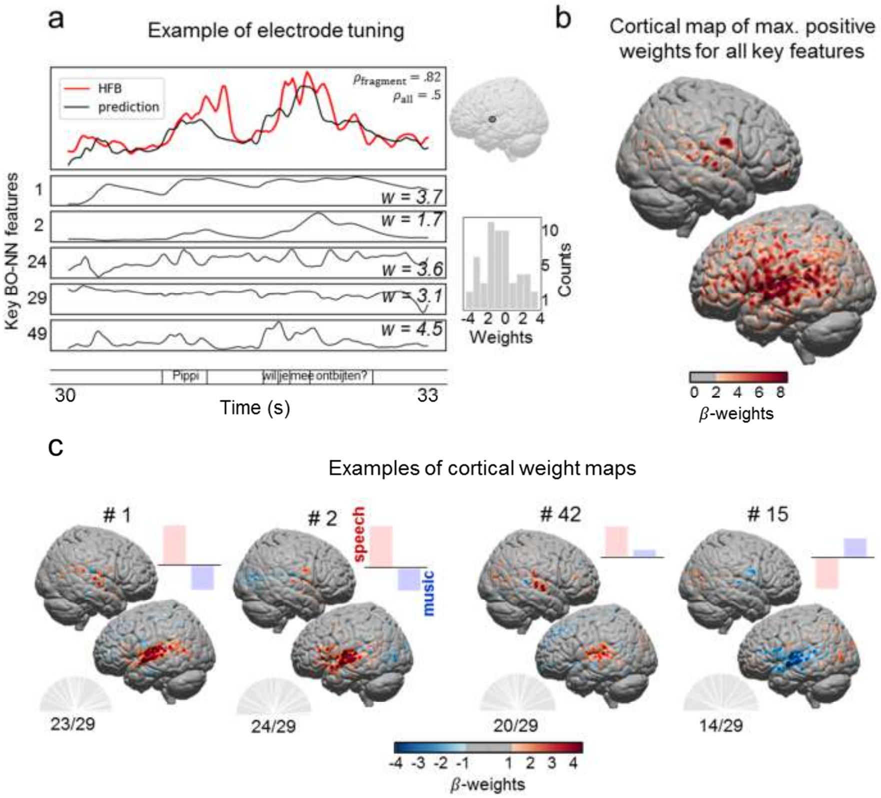


**Figure 3.5. Visualization and interpretation of key BO-NN features. (a)** Example of electrode tuning to various key features. Top right panel shows the location of the selected electrode on the MNI cortical surface. Bottom right panel shows distribution of weights over 53 clusters for the selected electrode. **(b)** Cortical map of maximal positive weights across all key BO-NN features. **(c)** Weight maps over electrodes for a number of key features. Per plot, bars in top right show normalized mean activity during speech vs music blocks, half-pie charts in bottom left show distribution of electrodes contributing to the corresponding cluster over all 29 subjects.

## Acoustic patterns captured in key BO-NN features

Next, we sought to interpret the key features through the acoustic properties they could have preserved. First, we observed that theory-driven STMFs were to a large extent captured in the key BO-NN features, based on the linear regression models fitted to predict STMFs from BO-NN features (**Figure 3.6a**).

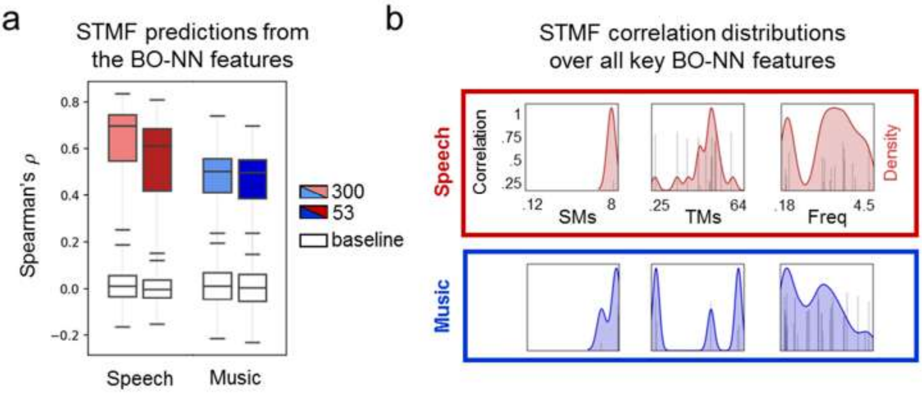


**Figure 3.6. Relation of key BO-NN features to STMFs. (a)** Prediction of STMFs from top BO-NN features, separately for speech and music, using either a full set of 300 features or only 53 key features. Prediction accuracy is estimated as a Spearman correlation ($\rho$) between predicted and observed SMTF feature time courses in a held-out test set. Shown is significant Spearman's $\rho$ (at $p < .001$) averaged over ten CV folds. **(b)** Distribution of highest Spearman correlations to STMFs over all key BO-NN features, separately for speech and music fragments. Dotted black lines show $\rho$ values per key feature (left axis), colored shaded areas display estimated kernel density over all features with significant correlation per STMF dimension: SMs, TMs or frequency (right axis).

To obtain more detail we examined significant Spearman correlations of each key BO-NN feature with all SMTFs, separately for speech and music (Movie II). In total, about a quarter of all key BO-NN features (14/52 in speech and 18/52 in music) showed large significant correlations ($\rho > .6$) with SMTFs. In case of speech, the dimension with most variance in correlation profiles across the key BO-NN features was TMs (**Figure 3.6b**). There was little variance along the dimension of SMs. The correlation to STMFs along the frequency dimension varied within 250 – 3000 Hz (typical range of formants).

In the case of music, the correlation profiles varied most along the frequency dimension (**Figure 3.6b**). There was more variance in correlations to SM features compared to speech but less variance in correlations to TMs compared to speech. Most features were correlated with STMFs in a low frequency range of 100 – 700 Hz and fine SMs in range of > 1 cycle/octave. Correlations to TMs were mostly limited to < 4 Hz.

In general, multiple key BO-NN features captured complex multidimensional acoustic profiles. The combinations across the TM-SM-frequency dimensions captured essential information characteristic of each type of sound. In music, we observed capturing of rich frequency-specific information, which changed slowly over time (slow TMs). In speech, we observed capturing of rich TM information in a range of frequencies typically associated with formants.

## Distinct temporal profiles of key BO-NN features

Finally, it has previously been shown that the neural responses to sound are characterized by multiple temporal and latency profiles (Berezutskaya et al., 2017b), therefore we investigated these in the key BO-NN features as well. First, we looked at the autocorrelation profiles of the key BO-NN features and found that some showed slow-wave responses whereas others showed faster temporal changes (**Figure 3.7a**). Most key features exhibited fast temporal profiles with autocorrelation dropping within 50 – 100 ms. Only a small set of key BO-NN features exhibited slow temporal profiles with high autocorrelation over 200 ms. These features were associated with cortical weight maps outside STG with coverage of frontal or MTG electrodes.

Focusing on perceived speech processing, we used basic speech properties, such as intensity, pitch, formant information and a binary vector of speech being present or absent in the sound (speechON) to predict the key feature activations during speech at different time lags (see Methods for details). Most key BO-NN features showed tuning to intensity, pitch and speechON at different time lags (**Figure 3.7b**). Best predictions were typically obtained at 100 – 150 ms of sound onset.

Some features did stand out for speech. For example, key feature #41 showed a peak response to speech intensity -50 – 0 ms prior to the onset. Key feature #1 was most tuned to speechON at a lag of 200 – 350 ms, and key feature #2 – at a lag of 250 – 300 ms. Some features were associated with negative weights for speech being on and sound intensity in general, and thus were “deactivated” by speech (key features #15 and #26). Notably, the autocorrelation profiles and the preferred latency were decoupled for many key features.

Overall, we observed a trend for an increased latency of response from -50 – 0 ms prior to speech onset (key feature #41) towards 300 ms after speech onset (key features #2 and #15, **Figure 3.7c**). Looking at the cortical maps of these features, we could also observe a trend for the electrode locations to move anterior along STG towards IFG. Notably, key BO-NN features #26 and #15 were largely “deactivated” by speech 300 – 400 ms after speech onset and were associated with a weight map over the dorsal LOTC.

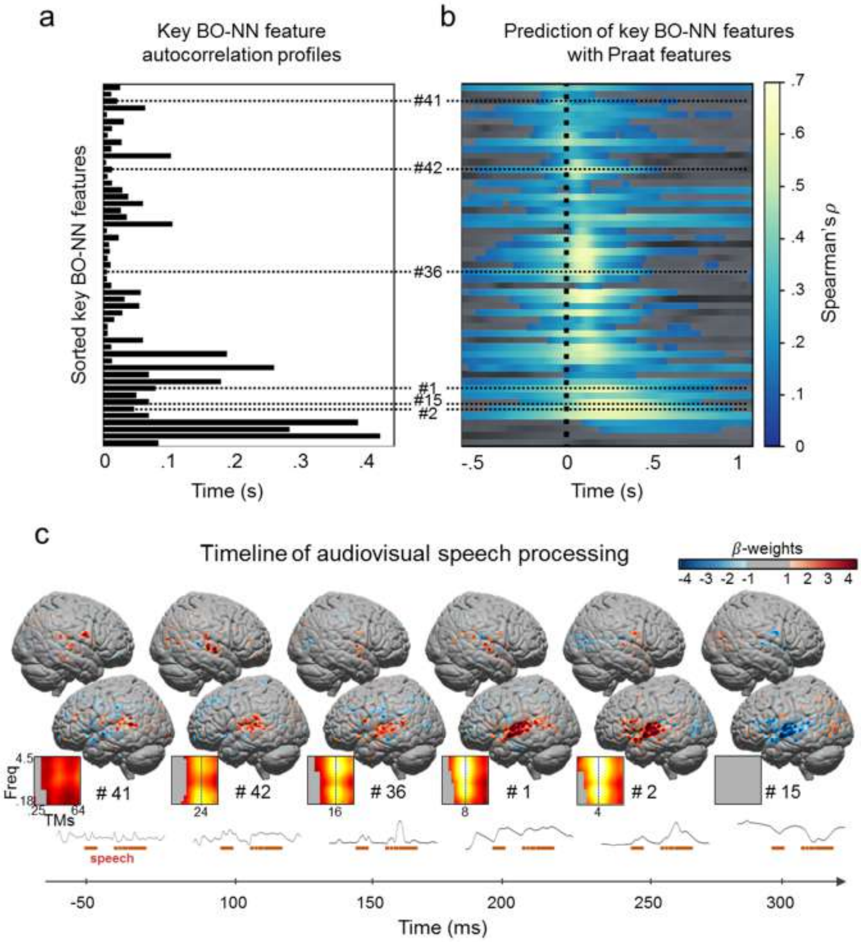


**Figure 3.7. Temporal profiles of the key BO-NN features. (a)** Autocorrelation profiles of 53 key BO-NN features. **(b)** Accuracy of predicting activity of the key BO-NN features from a set of speech features: speechON, sound intensity, pitch, first formant (F1) and second formant (F2) information. A linear model was trained per each audio-ECoG shift in range of -600 ms to 1 s. Spearman correlations between predicted and observed activity of the key BO-NN features were calculated and averaged over six cross-validation folds. **(c)** A tentative timeline of neural processing of audiovisual speech in -50 to 500 ms around the sound onset. The plots contain cortical weight maps for selected key features together with their example time course during a speech fragment. For each key feature, its STMF-correlation profile is shown for two dimensions: TMs and frequency. The dashed line indicated the maximal correlation along the TM dimension.

## 3.3 DISCUSSION

In the present study, we succeeded in constructing a model of auditory perception in the human brain by using a bottom-up approach while abstaining from injecting specific theoretical assumptions about the sound features. We trained a model on one set of data and subjects, and then validated it in new data and subjects. We found that a deep artificial neural network, trained on sound input to directly predict the associated neural responses, can provide such a model. This model was best at predicting the neural responses across lower-level and higher-level auditory processing cortices and it generalized well to unseen subjects and novel audio material. It uncovered a number of features, interpretable in terms of their acoustic, cortical and response latency profiles. Certain features were activated specifically by speech and reflected encoding of hierarchical temporal structures across the perisylvian cortex. Altogether, we argue that a data-driven approach with a minimum of theoretical assumptions provides a powerful model that can be used to test existing theoretical assumptions and generate new insights about the complex mechanisms of auditory perception in the brain.

Similarly to previous work (Güçlü et al., 2016; Kell et al., 2018), we observed that a hierarchical data-driven model, such as a deep artificial neural network (BO-NN) can produce high accuracy of predicting the neural responses to complex sensory input. This prediction accuracy exceeded accuracy obtained with a theory-driven STMF-based model (Berezutskaya et al., 2017b). Importantly, we were able to show that the performance of our data-driven model generalized well to a different experimental setup (different movie-watching experiment) and a large, different, set of participants.

The data-driven BO-NN model does not only lead to high accuracy of predicting brain data, but it also provides complex data-driven features, which lead to optimal predictions. In the present study we used feature visualization, clustering, correlation to known theory-driven features, and examination of cortical and temporal profiles, to gain insight into what information is captured in each data-driven feature. The features carry the potential to elucidate the nature of the relationship between perceived sound and the brain.

The power of the entirely data-driven approach is exemplified by the fact that the model extracted complex acoustic patterns that have been identified previously with a theory-driven model (SMs, TMs and frequency, Santoro et al., 2014; Schonwiesner and Zatorre, 2009). Some of the extracted features were sound-specific (either speech or music), suggesting that they could distinguish between types of sound, reflecting a higher-level kind of distinction. We have not yet observed a clear-cut distinction between low-level (acoustic properties) and higher-level (language or music types) features. The transformation from one to the other could be gradual and requires further examination of the features, perhaps along different layers of the deep neural network model.

Combining the information about the cortical distribution with the feature-specific acoustic and temporal latency profiles, we find agreement with a different theory of hierarchical processing of auditory input. In particular, the BO-NN model provides supporting evidence for hierarchical encoding of temporal information along the perisylvian cortex during speech perception (Honey et al., 2012; Hullett et al., 2016; Lerner et al., 2011; Poeppel, 2003), and for the notion of a posterior-anterior direction of temporal information propagation during naturalistic speech processing (**Figure 3.7c**, Berezutskaya et al., 2017; Hullett et al., 2016). Moreover, our results agree with the concept of multi-scale temporal processing of complex stimulus events in the human brain (Baldassano et al., 2017; Honey et al., 2012). Similar to previous reports on speech perception, we observed that neuronal populations in posterior STG encode fast TMs (> 10 Hz), whereas sites along the sylvian fissure in the anterior direction encode increasingly slower TMs up to the IFG (< 4 Hz).

Some of the present findings also suggest that the posterior STG area, surrounding the transverse temporal gyrus, may activate prior to, and at the moment of, speech onset. The effects of predictive coding (Friston, 2005) have previously been demonstrated in auditory perception, with similar spatiotemporal activity patterns for expected and perceived sounds (Bendixen et al., 2012; SanMiguel et al., 2013). Our results suggest presence of a neuronal component in posterior STG that may be active prior to speech onset in a continuous speech stream within a complex audiovisual narrative. Our finding provides additional support that the observed phenomenon is not necessarily due to the prediction violation mechanism, typically observed in studies with stimulus mismatch/omission paradigms (Bendixen et al., 2012), but more likely indicates anticipatory effects inherent to continuous speech perception.

Additionally, we observed involvement of the dorsal LOTC regions in the sound, in particular speech, processing network. This region responded to speech onset the latest (300 – 400 ms) and comprised a neuronal component that was deactivated by speech onset (key features #10, 15 and 26). These features did not correlate to acoustic sound properties and correlated negatively with the time courses across the perisylvian cortex. This anticorrelation decreased substantially in the absence of speech. This result can potentially be explained by the interaction of the visual and auditory domains in the movie, and a possible suppression of the visual motion information processing during speech. It has previously been shown that activation in this region can be modulated by attentional demands in the context of audiovisual stimulation (Cate et al., 2012).

Not all data-driven BO-NN features were characterized by a relation to acoustic properties of sound. Several key features were informative for prediction of signal in various cortical regions without a significant correlation to the STMFs. In addition, a number of features were characterized by slow-fluctuating time courses with a high autocorrelation rate and location in frontal brain regions. It remains to be further explored whether the extracted features can relate to other higher-level cognitive processes during movie watching, such as attention shifts, semantic information in the auditory stream and information of the movie plot present in both audio and visual modalities. In addition, more work is needed to explore the multi-layer audio representations, and not only the layer closest to the neural predictions (top BO-NN layer).

Overall, a combination of the neural responses during a naturalistic stimulus experiment with a bottom-up modeling approach allowed us to build a generalizable, top-performance neural model of auditory perception with a set of interpretable key features that capture multi-level sound properties. The current pretrained model may be directly applicable to predict ECoG responses in new datasets using either controlled or naturalistic stimuli. The BO-NN model can also be further fine-tuned to fit the neural responses from another experimental setup. For example, given a previously demonstrated positive relationship between HFB and fMRI responses (Haufe et al., 2018; Hermes et al., 2012), it can be of interest to apply this framework in the context of fMRI data and explore tuning in transverse temporal gyrus, primary visual areas and the medial parts of the human brain which are inaccessible to ECoG grids.

## Conclusions

In the present study, we constructed a data-driven neural model of auditory perception by training a deep learning model on raw audio input to directly predict the associated brain responses, and validated it in new data and subjects. Our brain-optimized model successfully fitted the brain responses along the perisylvian cortex. The extracted features captured complex spectrotemporal sound properties and exhibited distinct cortical distributions with specific latencies of response to sound onset. Several model features captured a temporal hierarchy of speech processing along the perisylvian cortex.

## 3.4 MATERIALS AND METHODS

### Movie stimuli

A Dutch feature film “Minoes” (2001) was used as stimulus for the first movie-watching experiment (Movie I). The film was 93 minutes long (78 minutes before credits) and told a story about a cat Minoes, that one day transforms into a woman. In her human form, she meets a journalist. Together, they solve several mysteries involving their town, and during their adventures eventually fall in love. The movie was made in Dutch and was easy to follow for both kids and adults. Almost all patients reported that they had enjoyed watching the movie.

In the second movie-watching experiment we used a 6.5 minute short movie, made of fragments from “Pippi Langkous” (Pippi Långstrump, 1969) edited together to form a coherent plot (Movie II). The task was part of the standard battery of clinical tasks performed with a purpose of presurgical functional language mapping. The movie consisted of 13 interleaved blocks of speech and music, 30 seconds each (seven blocks of music, six blocks of speech). The movie was originally in Swedish but dubbed in Dutch.

### ECoG experiments

All participants were admitted for diagnostic procedures with medication-resistant epilepsy. They underwent subdural electrode implantation to determine the source of seizures and test the possibility of surgical removal of the corresponding brain tissue for about a week. Research could be conducted between clinical procedures. All patients gave written informed consent to participate in accompanying electrocorticography (ECoG) recordings and gave permission to use their data for scientific research. The study was approved by the Medical Ethical Committee of the Utrecht University Medical Center in accordance with the Declaration of Helsinki (2013).

Eight patients (age 22±7, six females) watched Movie I. One patient was excluded from the analyses because no electrodes covered language regions (no significant response to language or audio tasks). Another patient was excluded due to severe and frequent seizures. Thus, data from six patients was used. Three patients were implanted with left hemispheric grids. Most patients had left hemisphere as language dominant, based on fMRI or the Wada test (**Table 3.1**).

Thirty-three patients (age 26±13, 21 females) watched Movie II. Of those, four were excluded from the analyses because those patients also participated in the experiment with Movie I. Of the remaining 29 patients, 23 were implanted with left hemispheric grids. Most patients had left hemisphere as language dominant, based on fMRI, Wada or functional transcranial Doppler sonography test (**Table 3.1**).

**Table 3.1. Electrode grid information for all participants (both Movie I and II).** Shown is information about the number of electrodes, grid hemisphere, covered cortices, handedness, and language-dominant hemisphere per patient. L, Left; R, right; F, frontal cortex; M, motor cortex; T, temporal cortex; P, parietal cortex; O, occipital cortex; fMRI, functional magnetic resonance imaging; fTCD, functional transcranial Doppler sonography.

| Patient | N of electrodes | Grid hemi-sphere | Cortices covered | Handed-ness | Language dominance | Movie experiment |
|---|---|---|---|---|---|---|
| 1 | 88 | R | M, T, P, O | R | L (fMRI) | I |
| 2 | 64 | R | F, M, T, P | R | L (fMRI) | I |
| 3 | 64 | R | T, P, O | R | L (Wada) | I |
| 4 | 64 | L | F, M, T, P | R | L (fMRI) | I |
| 5 | 64 | L | M, T, P, O | L | L (fMRI) | I |
| 6 | 120 | L | T, P, O | R | L (fMRI) | I |
| 7 | 96 | L | F, T, P | R | L (Wada) | II |
| 8 | 112 | L | F, M, T, P, O | R | L (fMRI) | II |
| 9 | 96 | L | F, M, T, P, O | R | L (Wada) | II |
| 10 | 104 | L | F, M, T, P | R | L (Wada) | II |
| 11 | 48 | L | F, M, T, P | L | L (fMRI) | II |
| 12 | 120 | L | F, M, T | R | L (Wada) | II |
| 13 | 112 | L | F, M, T, P | R | L (fMRI) | II |
| 14 | 64 | L | F, M, T | R | L (fMRI) | II |
| 15 | 112 | L | M, T, P, O | R | L (fMRI) | II |
| 16 | 64 | L | M, T, P | R | L (Wada) | II |
| 17 | 96 | R | M, T, P, O | L | R (Wada) | II |
| 18 | 88 | L | T, P, O | R | L (fMRI) | II |
| 19 | 112 | L | F, T, P | R | R (Wada) | II |
| 20 | 120 | R | F, M, T | R | L (Wada) | II |
| 21 | 96 | R | F, T, P, O | L | L (Wada) | II |
| 22 | 120 | L | F, T, P, O | not available | L (Wada) | II |
| 23 | 88 | L | F, M | R | L (fMRI) | II |
| 24 | 96 | L | F, M, T, P, O | L | L (Wada) | II |
| 25 | 64 | R | T, P, O | R | not assessed | II |
| 26 | 88 | L | F, M, P | R | L (fMRI) | II |
| 27 | 80 | L | F, M, T, P | R | L (Wada) | II |
| 28 | 128 | L | F, M, T, P, O | R | L (fMRI) | II |

| | | | | | | |
|---|---|---|---|---|---|---|
| 29 | 80 | R | F, M, T, P, O | R | L (fMRI) | II |
| 30 | 64 | L | F, M, T | R | L (fMRI) | II |
| 31 | 64 | L | F, M, T | R | L (fMRI) | II |
| 32 | 48 | R | F, M, T | L | R (fMRI) | II |
| 33 | 112 | L | F, M, T | R | R (fMRI) | II |
| 34 | 64 | L | F, M, T | R | L (fTCD) | II |
| 35 | 64 | L | F, M, T, P | R | L (fMRI) | II |

All patients were implanted with clinical electrode grids (2.3 mm exposed diameter, inter-electrode distance 10 mm, between 48 and 128 contact points); one patient (experiment with Movie II) had a high-density grid (1.3 mm exposed diameter, inter-electrode distance 3 mm). Almost all patients had perisylvian grid coverage and most had electrodes in frontal and motor cortices. The total brain coverage of the patients who watched Movie II can be seen in **Figure 3.2**. Patient-specific information about the grid hemisphere, number of electrodes, and cortices covered is summarized in **Table 3.1**.

In the movie-watching experiment (Movie I or Movie II), each patient was asked to attend to the movie displayed on a computer screen (21 inches in diagonal). The stereo sound was delivered through speakers with the volume level adjusted for each patient. Due to the long duration of Movie I, the patients were given an option to pause the movie and quit the experiment at any time. In that case, the patient could continue watching the movie at a later time starting from the frame they had paused on.

During both experiments, ECoG data were acquired with a 128-channel recording system (Micromed) at a sampling rate of 512 Hz filtered at 0.15–134.4 Hz. Both movies were presented using Presentation software (Version 18.0, Neurobehavioral Systems) and sound was synchronized with the ECoG recordings. Additional data, such as audio/video recordings of the room and PictureInPicture were used to ensure the synchronization.

## ECoG data processing

All electrodes with noisy or flat signal (visual inspection) were excluded from further analyses. After applying a notch filter for line noise (50 and 100 Hz), common average rereferencing was applied to all clinical grids per patient (and separately for the one high-density grid). Data were transformed to the frequency domain using Gabor wavelet decomposition at 1–120 Hz in 1 Hz bins with decreasing window length (4 wavelength full-width at half maximum). Finally, high frequency band (HFB) amplitude was obtained by averaging amplitudes for the 60–120 Hz bins and the resulting time series per electrode were down-sampled to 50 Hz. Electrode locations were coregistered to the anatomical MRI in native space using computer tomography scans (Branco et al., 2018; Hermes et al.,

2010) and FreeSurfer (http://surfer.nmr.mgh.harvard.edu/). The Desikan–Killiany atlas (Desikan et al., 2006) was used for anatomical labeling of electrodes (closest cortical structure in the radius of 5 mm). All electrode positions were projected to Montreal Neurological Institute space using SPM8 (Welcome Trust Centre for Neuroimaging, University College London).

### BO-NN model construction

We employed various artificial neural network architectures to train a model end-to-end by taking raw audio (sampling rate 16 or 8 kHz) as input and predicting the HFB neural activity as output. HFB time courses were concatenated across all electrodes of the six patients (396 electrodes in total). Specifically, we trained a convolutional neural network, a recurrent neural network and a recurrent convolutional neural network, each with either one (time) or two (time and frequency) input streams. Approximately 3,318,000 network parameters were tuned during the training. All networks were trained by minimizing the mean squared error using Adam optimization ($\alpha = 5 \times 10^{-4}$, $\beta_1 = .9$, $\beta_2 = .999$). The models were constructed and trained using a deep learning framework Chainer (Tokui et al., 2015).

The best performing model had a dual-stream architecture that combined both convolutional and recurrent layers. Similar result has been reported before for task-optimized artificial neural networks(Güçlü et al., 2016). The dual-stream model passed 1D raw audio waveform (six-second input fragment at 8 kHz) through five 1D convolutional layers and 2D audio spectrogram (six-second input fragment at 50 Hz) – through four 2D convolutional layers, concatenated the output and passed it to the recurrent layer with long-short-term-memory (LSTM) cells. Each convolutional layer was followed by a batch normalization layer, rectified linear activation function and local pooling.

### BO-NN model performance with Movie I (six participants)

Model performance was quantified as the Spearman correlation between predicted and observed HFB responses in the test set of Movie I (10% of all data):

$$\rho = \frac{\mathrm{cov}(rx,\ ry)}{\sigma_{rx}\sigma_{ry}} \tag{3.1}$$

where $rx$ and $ry$ are rank-transformations of $x$ and $y$, respectively, and $\sigma_{rx}$ and $\sigma_{ry}$ are the standard deviations of the rank variables.

A ten-fold cross-validation was used for an unbiased estimation of the model performance. Per each test fold the training continued for 30 epochs (i.e. full runs through all training data). The model estimated at the epoch with the best test performance across all test folds was used later on for testing of performance generalization to a dataset of Movie II.

The significance of the prediction accuracy was assessed parametrically by transforming Spearman's $\rho$ scores to $t$-values (equation 2) and determining the $p$-values based on the probability density of the Student's distribution. The reported $\rho$ scores were significant at $p < 1 \times 10^{-10}$, Bonferroni corrected for the number of electrodes and test folds.

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}} \tag{3.2}$$

To assess whether there was any difference in model performance depending on the type of the predicted sound, we extracted and separated speech, music, noise or ambient sound fragments from each test fold based on the manual annotation of the soundtrack. The fragments of one type of sound were concatenated within each fold. The fragment length across different types of sound was balanced by subselecting fragments of 200 s per sound type and per fold. The prediction accuracy was calculated separately for each fold and averaged across the folds. Significance testing was performed in the same way as described above.

### BO-NN model performance with unseen data (29 new participants, Movie II)

The soundtrack of Movie II was passed through the pretrained BO-NN model to obtain the feature activations for a new dataset. The model weights were frozen and the model was not fine-tuned to optimize the predictions on the dataset of Movie II. Then, a ridge linear regression (equation 3) was trained on each layer $l$ of the BO-NN model to predict HFB responses (matrix $\mathbf{Y}$ of size $N_{timepoints} \times N_{electrodes}$) to the sound features $\mathbf{X}_l$ of Movie II:

$$\mathbf{Y}_l = \mathbf{B}_l{}^{\mathrm{T}}\mathbf{X}_l + \boldsymbol{\varepsilon}_l \tag{3.3}$$

where $\boldsymbol{\varepsilon}_l \sim \mathcal{N}(0, \sigma^2)$

The optimal value of the regularization parameter $\lambda$ (equation 4) was determined through a nested ten-fold cross-validation. All linear models were fitted using a machine learning Python package scikit-learn (Pedregosa et al., 2011).

$$\widehat{\mathbf{B}_l} = \underset{\mathbf{B}\in\mathbb{R}}{\operatorname{argmin}} \left\| \mathbf{Y}_l - {\mathbf{B}_l}^{\mathrm{T}}\mathbf{X}_l \right\|_2^2 + \lambda \left\| {\mathbf{B}_l}^{\mathrm{T}} \right\|_2^2 \qquad (3.4)$$

Because the dataset of Movie II comprised interleaved blocks of speech and music, we trained separate ridge regression models to predict HFB responses to speech and music separately. As before, the model performance was measured as a Spearman correlation between the predicted and the observed HFB responses in the test set. The model performance was cross-validated across six 30-second test folds. Thus, each sound block of speech or music was used as a test fold to minimize the effect of variance across the blocks.

The significance of the linear model performance was assessed using an exact test. For each model (each layer of the BO-NN model and separately for speech and music) we shifted the time courses of the HFB responses to disrupt the alignment with the sound and trained a ridge regression model using the shifted time courses. This resulted in a baseline distribution of accuracy scores (n = 1,000). The significance threshold for the performance of the models with aligned time courses was set at the 99.999th percentile of the baseline, shifted, distribution, which corresponds to a $p$-value of $.001$.

The performance across layers of the BO-NN model was compared using one-sided Wilcoxon signed tests. The layer comprising LSTM cell states showed highest prediction accuracy and was used in all further analyses. It is referred to as the top layer of the BO-NN model. The hidden states of the LSTM layer showed similar prediction accuracy.

## Comparison of the data-driven BO-NN features with the theory-driven feature set

As a state-of-the-art theory-driven feature set we used STMFs (Chi et al., 2005), that was previously used with this dataset(Berezutskaya et al., 2017b). The features of the modulation set were constructed based on a biological model of the cortical processing of the time-frequency representation of sound. STMFs were defined along the dimensions of temporal modulations (TMs) and spectral modulations (SMs), for both time and frequency dimensions of the spectrogram. The features were extracted by convolving the time-frequency representation of sound (cochleogram) with 2D filters with predefined sizes and orientations. Based on the previous work (Berezutskaya et al., 2017b; Kell et al., 2018; Pasley et al., 2012), here we used a range of .125 – 8 cyc/oct for SMs and a range of .25 – 64 for TMs.

To ensure a fair comparison with top BO-NN, we selected the best performing set of STMFs by varying the size of the filters along both temporal and spectral dimensions as well as subsampling along the frequency axis. In total, we looked at eight various subsets comprising 63, 119, 221, 252, 936, 952, 1768 or 3536 features. The subsets included reduced ranges of SMs (only $> 1$ cyc/oct) and TMs (only $\leq 4$ Hz), and subsampled or averaged frequency bins (1, 4, 8 or 16). The feature set selected for the comparisons contained 13 SMs in range of .125 – 8 cyc/oct, 14 TMs in range of .25 – 64 and eight frequency bins (from averaging instead of subsampling). This feature set achieved one of the highest accuracy scores in predicting the HFB neural responses.

To compare the representational geometries across different models and identify the feature set with the representational geometry most similar to the neural responses in STG we employed the representational similarity analysis (RSA, Kriegeskorte et al., 2008; Nili et al., 2014). Within the RSA framework, internal representations of various candidate feature sets (top BO-NN and STMFs) are compared to the target dataset (neural responses in STG) to assess which representation is least dissimilar to the target dataset. Here, we used Spearman correlation ($\rho$) across $N$ time points ($N \times N$ matrices, vectorized) as the internal representation of each dataset ($N \times F$, where $F$ is the feature dimension). This was calculated separately for the speech and music fragments of Movie II. The dissimilarity between the target and the candidate datasets was calculated as $1 - \rho$ of their internal representations.

The RSA was performed separately for the speech and music fragments to account for the possible difference in representation of different types of sound. All sound-specific fragments were concatenated in time and split into 6-second chunks, resulting in 30 speech and 30 music chunks. Then, per chunk, the first order dissimilarity matrices were constructed by computing $1 - \rho$ (Spearman correlation) across the 6 seconds by correlating feature vectors in each feature set (candidate datasets) and the neural responses in STG electrodes (target dataset) across time points of the chunk. Thus, all STG electrodes (199 in total) were used as features of the STG representation. Then, per chunk, the second order dissimilarity matrix was constructed by vectorizing the first order dissimilarity matrices per feature set, as well as for the target dataset, and then computing $1 - \rho$ between the vectors. The resulting dissimilarity values ($1 - \rho$) were compared between the target dataset and each of the candidate dataset in all 30 sound-specific chunks for both speech and music. The closest match to the target dataset was determined via a non-parametric two-sided Wilcoxon signed test using a dissimilarity value per each chunk as a single data point.

To compare performance across the feature sets in higher-level cortical areas during speech we trained a ridge linear model to fit each electrode in STG, IFG, postcentral gyrus and MTG. The linear model training and evaluation were performed in the same way as described previously, by employing a nested cross-validation to estimate the regularization parameter and computing prediction accuracy by cross-validating per-block performance. The cross-validated performance was compared between the three models in a non-parametric two-sided Wilcoxon signed test per ROI using each electrode's performance as a single data point.

### Extraction of key BO-NN features

To minimize the number of features to work with, we identified the key features of top BO-NN using another data-driven approach – applied affinity propagation (AP) clustering (Frey and Dueck, 2007). This approach offers a benefit of identifying key data points (cluster centers) that are most representative of the clusters they belong to. This means that there is no additional transformation from the original features to a new feature space. This is convenient because we can directly visualize the cortical weight map associated with each cluster center (key feature), as it is simply the weight map of the feature that already exists in the dataset. The optimal number of clusters was identified in a data-driven way by varying the parameter of *preference* that affects the a priori suitability of each data point to be a cluster center. As the number of clusters changes depending on the *preference*, the optimal estimate will fall in the knee of the curve, representing a balance between maximum compression and optimal cluster assignment accuracy.

AP clustering was performed on the electrode $\beta$-weight profiles, which were output in a ridge linear regression that predicted HFB responses from the 300 top BO-NN features. The $\beta$-weights were averaged over the six cross-validation folds. We took the $\beta$-weights from the linear model that was tested on the speech fragments. Another regression was trained to predict the music fragments. Both models were exposed to roughly the same amount and type of data in the training set, and the $\beta$-weights did not differ significantly between them.

AP clustering was applied to the $\beta$-weight profiles using the Spearman correlation matrix across the weights as the *affinity* matrix. This clustering approach was chosen because it estimates the optimal number of clusters in a data-driven way. Specifically, this was done by varying the *preference* parameter to determine the optimal number of clusters (53 key BO-NN features) which corresponded to the knee of the curve.

Additionally, we estimated the accuracy drop-off depending on the number of clusters separately in speech and music by using only the time courses of the cluster centers (different depending on the preference parameter) as predictors of the HFB responses. This was done to check that reduction of the key BO-NN features did not dramatically affect the accuracy of predicting the neural responses in Movie II.

### Selection of speech-specific and music-specific features

Features with peak activation during speech and music fragments were identified by comparing the mean normalized activation across the fragments. In case of speech, we used manual annotations for word onsets and offsets rather than the entire speech block because of the presence of a large number of pauses in speech, whereas the music was constantly present in the music blocks.

In order to test feature specificity to speech or music we split all data into 'trials' of 6 s, resulting in 30 trials of speech condition (six speech blocks, 30 s each) and 30 trials of music condition (seven music blocks, 30 s each minus one block that repeated the music content of a previous block). We averaged the feature activation over time per each 'trial' separately for speech and music trials. Then we assessed the difference in average activation amplitude between speech and music trials separately for each feature using non-parametric two-sided Wilcoxon signed-rank tests.

### Acoustic patterns captured in key BO-NN features

A ridge linear regression was used to predict modulation features from the full set of the top-layer BO-NN features as well as a set of 53 key BO-NN features. Separate regression models were trained to predict speech and music fragments. The regularization parameter was determined through the nested cross-validation. The prediction accuracy scores were cross-validated across the six sound-specific blocks. Significance testing was in the same way as before by creating a baseline distribution of prediction accuracy using shifted time courses of modulation features ($n = 1{,}000$). Similar to the previous analyses, the significance threshold was set at the 99.999th percentile of the baseline distribution, which corresponds to a $p$-value of $.001$.

Feature-to-feature Spearman correlations between the key BO-NN features and the spectrotemporal modulation features were computed separately for speech and music fragments. The data across all sound-specific blocks was concatenated into one vector and correlated between the feature sets. Significance testing was based on the correlations between the key BO-NN features and the shifted modulation feature time courses ($n = 1{,}000$, shifts > 6 s). The $p$-value threshold was set at the 99.999th percentile of the baseline correlation distribution.

The maps of correlation distributions over all key BO-NN features were created by selecting features with significant correlation to STMFs, identifying the combination of values along dimensions of SMs, TMs and frequency, associated with the highest correlation and calculating the variance in correlation along each STMF dimension. The key features were then separated into groups depending on the dimension of the highest variance in correlation. The procedure was carried out separately for correlations during music and speech fragments.

## Temporal profiles of the key BO-NN features

Autocorrelations were calculated for each of the key BO-NN features using the entire feature time course. The autocorrelation value was determined as a time point as which the autocorrelation dropped below $r = .5$.

Latency profiles were calculated only for the speech fragments. First, we trained a ridge linear regression to predict the time courses of the key BO-NN features based on the simple speech audio characteristics, such as intensity, pitch, formant values and a binary regressor of speech being on or off (based on the manual annotation of the speech fragments). The continuous speech regressors (intensity, pitch and formants) were extracted automatically from the soundtrack using Praat software (Boersma and Weenink, 2016). The binary regressor (speechON) was labeled manually by annotating the speech fragments with onsets and offsets of words. All the time points between consecutive onsets and offsets were labeled as ones and remaining time points were labeled as zeros. The key feature activations were then modeled as a linear combination of all the simple speech regressors. The details of training and cross-validation of the ridge linear regression were identical to the ones described above with the exception of additional training of separate models for different shifts between the speech regressors and key BO-NN feature activations to account for a possible temporal lag in the BO-NN feature activation to the speech regressor. A separate model was trained per each shift within -.5 to 1 second. The significance testing was performed in the same way as previously described by using shifted time courses and constructing a baseline distribution of the performance accuracy. In this case, the shifts were larger than 6 seconds. The latency per key BO-NN feature were determined as the maximal prediction accuracy, falling above the 99.999th percentile of the baseline correlation distribution, similarly to the previous analyses using the baseline shifted distribution to determine significance.

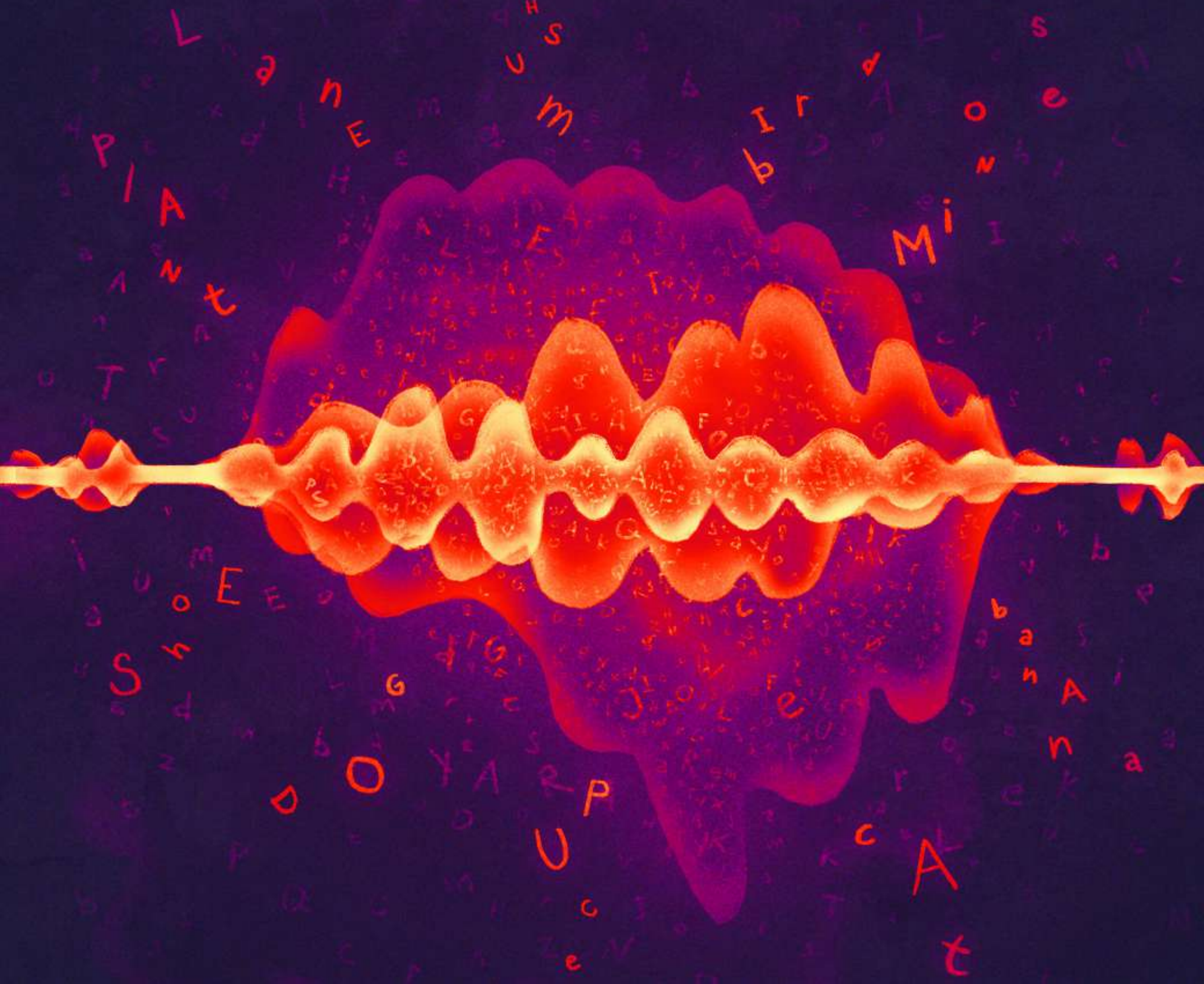
## Data and code availability

The data supporting the current study have not been deposited in a public repository due to the restrictions on public sharing of the patients' data but are available from the corresponding author on request. The code containing the BO-NN architecture, the details of training and the weights of a trained model are available in our online repository: https://github.com/Immiora/bonn_auditory_perception_ecog.

chapter

4

# High-density intracranial recordings reveal a distinct site in anterior dorsal precentral cortex that tracks perceived speech



# CHAPTER 4. HIGH-DENSITY INTRACRANIAL RECORDINGS REVEAL A DISTINCT SITE IN ANTERIOR DORSAL PRECENTRAL CORTEX THAT TRACKS PERCEIVED SPEECH[3]

Various brain regions are implicated in speech processing, and the specific function of some of these regions is better understood compared to the function of others. In particular, the involvement of the dorsal precentral cortex (dPCC) in speech perception remains debated, and attribution of the function of the region is more or less restricted to motor processing. In this study we investigated high-density intracranial responses to speech fragments of a feature film, aiming to determine whether dPCC is engaged in perception of continuous speech. Our findings show that dPCC exhibited preference to speech over other sounds. Moreover, the identified area was involved in tracking speech auditory properties including speech spectral envelope, its rhythmic phrasal pattern and pitch. DPCC also showed the ability to filter out noise from the perceived speech. Comparing these results to data from motor experiments showed that the identified region had a distinct location in dPCC, anterior to the hand motor area and superior to the mouth articulator region. The present findings uncovered with high-density intracranial recordings help elucidate the functional specialization of PCC and demonstrate the unique role of its anterior dorsal region in continuous speech perception.

[3] This chapter is based on Berezutskaya, J., Baratin C., Freudenburg, Z. V. & Ramsey, N. F. High-density intracranial recordings reveal a distinct site in anterior dorsal precentral cortex that tracks perceived speech. *In review.*

## 4.1 INTRODUCTION

It is widely known that speech perception engages a large network of brain regions. The involvement and functional role of some of these regions is better understood compared to other regions. In particular, various theories implicate importance of superior temporal, middle temporal and inferior frontal gyri in language processing (Friederici, 2012; Hagoort, 2013; Hickok and Poeppel, 2007). Activation of these regions has been extensively demonstrated for audiovisual and purely auditory speech perception (Crinion et al., 2003; Wilson et al., 2008), for perceiving intelligible or noisy speech (Scott et al., 2000) and has been related to both semantic and syntactic processing (Rogalsky and Hickok, 2009).

When it comes to the sensorimotor involvement in speech processing, regions within the precentral cortex (PCC) have also been shown to engage during both speech production and speech perception (Pulvermüller et al., 2006; Skipper et al., 2007; D'Ausilio et al., 2009; Cheung et al., 2016; Skipper et al., 2017). The area of PCC whose language function is more comprehensively described is the ventral portion of PCC. This region is otherwise known as the 'face area' that controls facial and articulator movements necessary for speech production. Neuroimaging and electrophysiological studies have yielded detailed maps of mouth articulators in ventral PCC (Bleichner et al., 2015; Bouchard et al., 2013; Chartier et al., 2018). As the region also shows reliable activation during speech perception (Murakami et al., 2011; Skipper et al., 2005; Watkins et al., 2003), a number of theories have been proposed to explain how the primary function of the 'face area' in speech production drives its responses to speech perception (Galantucci et al., 2006, Liberman and Mattingly, 1985; Skipper et al., 2006; Hickok and Poeppel, 2007, see Skipper et al., 2017 for a review).

Recent studies have also implicated a more dorsal region of the PCC, an area adjacent to the region associated with upper limb motor control (Begliomini et al., 2008; Bleichner et al., 2015; Roland et al., 1980; Schellekens et al., 2018) which is often referred to as the 'hand knob' (Yousry et al., 1997). Another adjacent region within dorsal PCC region has been associated with the motor function of larynx (Dichter et al., 2018; Simonyan and Horwitz, 2011) and with production of speech (Bouchard et al., 2013; Brown et al., 2007; Dichter et al., 2018; Olthoff et al., 2008), singing (Dichter et al., 2018) and vocalization in general (Brown et al., 2009).

In spite of its apparent specialization in motor control, dPCC and the neighboring cortex have been implicated in speech perception as well (Floel et al., 2003; Glanz et al., 2018; Keitel et al., 2018; Wilson et al., 2004). Some studies explain the activation of dPCC during speech perception through feedforward articulation-to-audio predictions (Meister et al., 2007). Some other work points towards the facilitation function of dPCC in perception of speech under difficult conditions, where its activation compensates for noisy input to the auditory cortex and aids in discrimination of speech sounds (Du et al., 2014; Wilson and Iacoboni, 2006). Another line of research connects the involvement of dPCC in speech perception to cortical entrainment of rhythm, phrasal speech rates and encoding of the temporal structures in perceived stimuli (Bengtsson et al., 2009; Keitel et al., 2018). It is important to note that these studies focus on neural activation to perceived speech in general, rather than responses to individual words or phrases semantically related to hand, face or body actions (see the theory of grounded cognition by Barsalou, 1999; Barsalou et al., 2003 and many related works, for example Hauk et al., 2004; Raposo et al., 2009; Shtyrov et al., 2014).

Given that it is unclear how a region typically associated with motor planning and execution can be involved in speech perception, we here seek to elucidate such involvement. We report results of a rare opportunity to investigate the role of dPCC in speech perception from high-density (HD) electrode grids placed in patients with epilepsy. These grids provide a unique combination of high temporal and spatial resolution that offers high detail of the underlying brain function (Jerbi et al., 2009). High-density (HD) recordings obtained directly from the cortical surface preserve information often underrepresented or lost in other neuroimaging modalities (Berezutskaya et al., 2017b; Dalal et al., 2009). Rather than employing specific tasks, the choice of which restricts evoked neural responses to constrained cognitive concepts (Brennan, 2016; Chen et al., 2013; Schmidt et al., 2008), we investigated data obtained while participants engaged in watching a full-length film. Such a naturalistic approach has been reported to be particularly beneficial in assessing cortical representation of a complex cognitive function such as speech processing (Glanz et al., 2018; Hamilton and Huth, 2018; Honey et al., 2012).

Data were collected from two patients implanted with HD grids (3 or 4 mm inter-electrode distance) and two with the standard clinical grids (10 mm inter-electrode distance). The HD electrodes were placed over the PCC. The neural responses to speech and non-speech film fragments were analyzed.

We found that dPCC showed increased responses to speech compared to other auditory input. We were able to show that dPCC had the capacity to filter out background noise from the perceived speech signal and tracked various auditory properties of speech, such as its rhythmic phrasal structure, spectral envelope and pitch. None of these auditory properties were tracked in the non-speech input. Importantly, we demonstrate that the location of the identified region is different from the hand motor and mouth articulator regions. The observed effects were prominent in HD data but substantially weaker in the participants implanted with clinical intracranial grids. These results underline the specific function of dPCC in tracking of perceived speech and have direct implications for our understanding of the neural processes underlying continuous naturalistic speech perception.

## 4.2 RESULTS

In this study, we investigated the involvement of dorsal precentral cortex (dPCC) in naturalistic speech perception using high-density (HD) intracranial electrode recordings. Two participants (S1 and S2) implanted with HD grids over the PCC watched a full-length feature film (**Figure 4.1a**). We then analyzed their brain responses in 65-125 Hz (high frequency band, HFB) (Crone et al., 1998; Ray et al., 2008) in relation to the speech fragments of the film. First, in each subject we identified a set of electrodes in dPCC with significantly higher HFB responses to speech compared to non-speech fragments (music, noises, animal cries, etc.). Then, we investigated the relations between the responses of these electrodes and various auditory properties of speech, such as the speech spectral envelope (associated with loudness, pitch, timbre and rhythm), its rhythmic phrasal pattern and pitch. We found a significant amount of neural tracking of these auditory properties. We also examined neural tracking of noisy speech fragments and found that dPCC electrodes had the ability to filter out background noise during perception of speech. Interestingly, the effects reported here were strong in participants with HD electrode grids, but were substantially less clear in participants with low-density electrode grids.

### Preference to speech fragments in dPCC

First, we aimed to determine whether any parts of PCC showed larger response amplitude during speech perception compared to perception of other sounds. Because PCC is not generally considered as part of the sound processing network, we did not additionally evaluate whether it generally exhibited a higher response to sound compared to silence. Our goal was to determine speech-specific response in PCC, thus we considered non-speech sounds as a baseline for our comparison. Thus, for our speech/non-speech comparison, we extracted 115 four-second-long fragments of each group (speech and non-speech fragments) and compared the average HFB responses associated with each group. The non-speech fragments contained various sounds, such as music, environmental noises (e.g. thunder, animal cries etc.), technical noises (e.g. car noises, phone ringing etc.), footsteps, clapping, etc. HFB responses were averaged per each fragment, and the groups were compared using independent samples t-tests (see Methods for details).

The t-tests between speech and non-speech fragments were conducted per electrode. They showed that 20 electrodes in S1 and 41 electrodes in S2 on average exhibited higher responses to speech: $t_{S1}$ranged from $3.9$ to $27.85$ and $t_{S2}$ ranged from $4.91$ to $27.74$ (df $= 228$) at $p < .05$, Bonferroni corrected for the total number of electrodes (Bonf. cor., **Figure 4.1b**). The reported ranges include only significant electrodes, which were found anterior to precentral sulcus in both subjects and corresponded to dPCC. The electrode locations formed consistent clusters in both subjects. Two more electrodes with significant effect in S1 were located in the superior temporal cortex (STG) and four more electrodes with significant effects in S2 were located in ventral PCC. We used the outcome of this analysis (significant t-values) to restrict the number of electrodes used in further analyses to 20 electrodes in S1 and 41 electrodes in S2.

Thus, we have observed that a subset of electrodes in dPCC showed preference to speech over other auditory input. To understand the nature of the dPCC response to speech better, we investigated its activity in relation to various perceptually relevant properties of speech. First, we examined whether dPCC tracked the overall shape, or envelope, of the spectrotemporal speech signal that is relevant for perception of consonants and vowels, and overall speech intelligibility. Then, we investigated dPCC responses to speech in noisy conditions, which is relevant for perception of speech in mixed auditory input. Third, we examined the neural activity related to the rhythmic structure of speech that is relevant for parsing of the continuous speech input into meaningful groups. Finally, we investigated the encoding of pitch, which is relevant for perception of intonation changes and speaker identification.

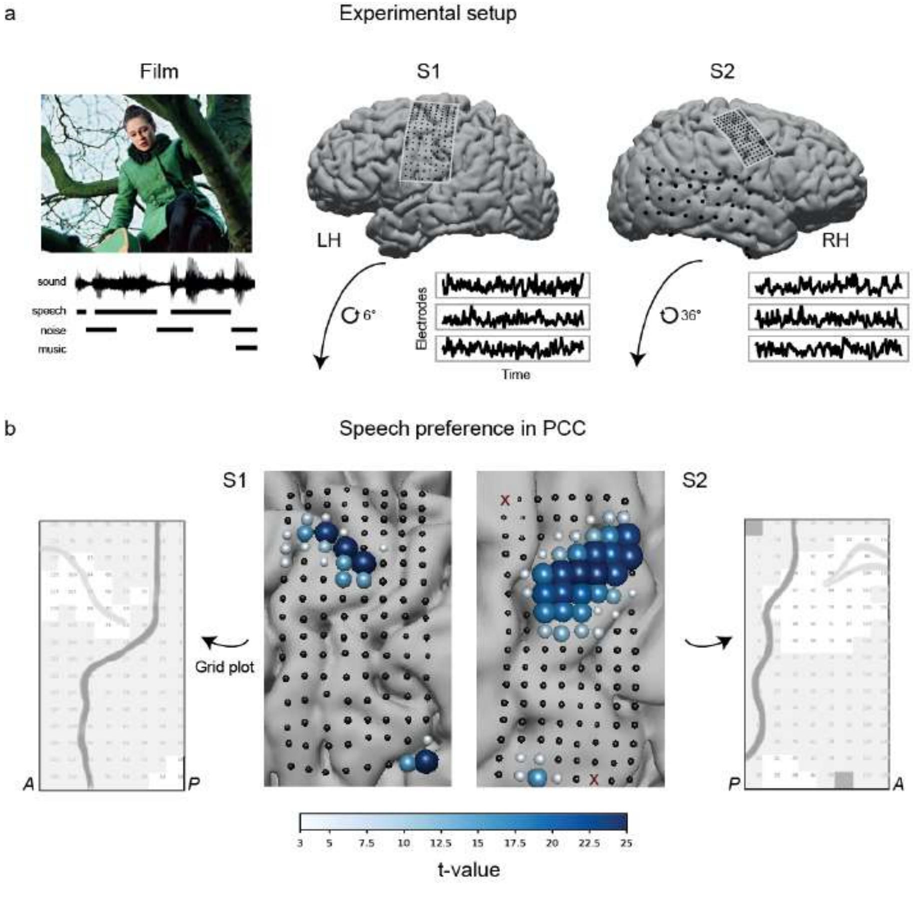


**Figure 4.1. (a)** Experimental setup. Two participants watched a full-length feature film Minoes (2001). The soundtrack was annotated with fragments of speech, noise and music sounds. During the experiment the neural responses were collected with HD ECoG grids placed over the sensorimotor cortex. A standard-density grid in S2 recorded from the temporal lobe. Recordings from a standard-density grid in S1 (not shown) were contaminated with epileptic seizures and were not analyzed. Example HFB time courses in three electrodes are shown per subject. **(b)** Results of the t-tests comparing average HFB activity during speech and non-speech (noise, music) fragments. HD grids were rotated for visualization. Each point on the grid is an ECoG electrode. Size and color of each electrode represents the t-value. Positive t-value represents higher average HFB activity during speech (p < .05, Bonferroni corrected for number of electrodes). Next to the brain plots displayed are the schematic grid plots, which are used for visualization in further analyses. Each square refers to one electrode. The numbers correspond to electrode indices. Darker grey line shows the outline of the precentral sulcus. Lighter grey line shows the outline of the superior frontal sulcus. Greyed out electrodes in S2 show electrodes containing artifacts. These electrodes were excluded from all the analyses. *A*: anterior direction (towards the frontal lobe), *P*: posterior direction (towards the occipital lobe).

## Tracking of speech spectral envelope in dPCC

### Correlation to spectral envelope of speech

To begin with, we focused on a slow-varying spectrotemporal feature of the speech sound, the spectral envelope. The spectral envelope is computed from the speech signal transformed to the time-frequency domain. It is computed for each time point as the signal energy averaged over all frequencies relevant to speech (180 – 7200 Hz, Chi et al., 2005). It captures perceptually relevant characteristics of consonants and vowels and the temporal structure of speech (Ter Keurs et al., 1992), preserves spectral information reflecting speaker identity (Carey et al., 1996; Kitamura and Akagi, 1995) and is important for overall speech intelligibility (Arai et al., 1996; Ter Keurs et al., 1993). Thus, we tested whether dPCC electrodes responded to the speech spectral envelope by cross-correlating it to the HFB responses in the electrodes that displayed significant effects in the previous t-test comparing responses to speech and non-speech fragments. Spearman cross-correlation ($\rho$) was performed per each speech fragment (n = 115) and compared to the control condition (non-speech fragments) to test whether tracking of spectral envelope was stronger in speech than non-speech fragments.

A subset of previously defined electrodes (from the t-test above: 20 in S1 and 41 in S2) showed higher correlation to the speech spectral envelope ($\bar{\rho}_{\text{S1}} = .25 \pm .04$ and $\bar{\rho}_{\text{S2}} = .27 \pm .03$) compared to non-speech spectral envelope ($\bar{\rho}_{\text{S1}} = .16 \pm .01$ and $\bar{\rho}_{\text{S2}} = .17 \pm .03$), as indicated by t-tests (on the Fisher-transformed $\rho$ -values) at $p < .01$, Bonf. cor. (**Figure 4.2a**). The reported values show mean $\rho$-values and standard deviations over all significant electrodes ($\rho$-values were first averaged over all speech or non-speech fragments per electrode). The effect was significant for 7 out of 20 electrodes in S1 and 13 out of 41 electrodes in S2. All electrodes that showed significant tracking of the speech spectral envelope (compared to non-speech baseline) were localized to the dorsal portion of precentral gyrus, except for two STG electrodes in S1.

In addition, we observed that the highest $\rho$-values typically fell in the range of 200 – 400 ms after sound onset (**Figure 4.2a**, bottom panel) suggesting that there was a positive ≈300 ms lag of speech tracking in dPCC.

## Correlation to STG electrodes

To further investigate the neural tracking of speech spectral envelope in dPCC, we assessed the relationship between dPCC and electrodes directly involved in auditory processing, such as STG electrodes. An increased correlation between dPCC and STG during speech perception could be due either to elevated communication between the two regions during speech perception or independent involvement in processing of speech fragments. These can be distinguished by examining the lag of correlation as that can reveal a latency between the regions. A lag would indicate dependency, while no lag would indicate that both regions process speech input in parallel to each other. Spearman cross-correlation scores between dPCC and STG electrodes were calculated during speech fragments and compared to the cross-correlation scores in non-speech fragments with t-tests. For this, in each subject we first identified a single STG electrode with strongest speech tracking (i.e. highest correlation to speech envelope: $\bar{\rho}_{\mathrm{S1}} = .34 \pm .14$; $\bar{\rho}_{\mathrm{S2}} = .32 \pm .16$) and then cross-correlated its time course with all previously selected dPCC electrodes.

T-tests between Fisher-transformed dPCC-STG correlations in speech and non-speech fragments resulted in a set of dPCC electrodes with higher correlation to STG during speech ($\bar{\rho}_{\mathrm{S1}} = .29 \pm .02$ and $\bar{\rho}_{\mathrm{S2}} = .35 \pm .02$) compared to the non-speech condition ($\bar{\rho}_{\mathrm{S1}} = .22 \pm .02$ and $\bar{\rho}_{\mathrm{S2}} = .25 \pm .03$, **Figure 4.2b**). The reported values show mean $\rho$-values and standard deviation over all significant electrodes ($\rho$-values were first averaged over all speech or non-speech fragments per electrode). The location of these electrodes was similar to the previous analysis (cross-correlation to speech envelope) and restricted to dorsal precentral gyrus (except for STG electrode 32 in S1). The electrode set included 5 out of 20 electrodes in S1 and 5 out of 41 electrodes in S2. The lag of maximal dPCC-STG correlation varied across the fragments, fluctuating mostly around zero, which indicated that both regions likely tracked speech input parallel to each other.

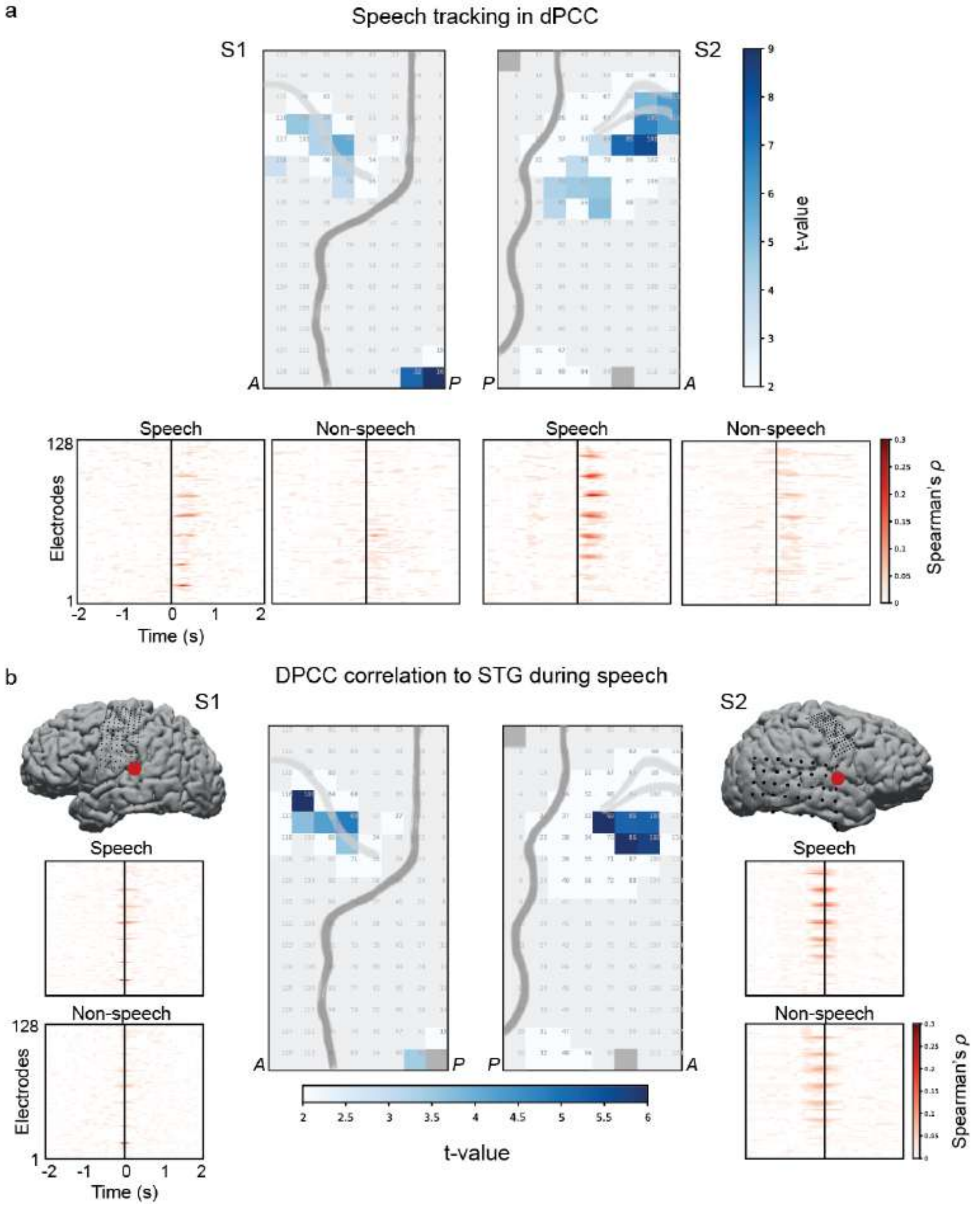


**Figure 4.2. Speech tracking in dPCC. (a)** Top panel shows results of the t-tests comparing the amount of correlation to the audio spectral envelope in speech and non-speech fragments (p < .01, Bonferroni corrected for number of electrodes). Bottom panel shown individual cross-correlation profiles per electrode separately for speech and non-speech fragments. The plots are averaged over the fragments (115 speech and 115 non-speech). For the t-tests the maximal correlation was taken in a range of -100 to 500 ms per electrode (one maximal correlation per fragment of each condition). The correlation

values were Fisher-transformed and fed to the independent two sample t-test per electrode. The results were Bonferroni corrected for the number of t-tests (=number of electrodes). **(b)** Shown in center are the results of the t-tests comparing the amount of correlation to a STG electrode in speech and non-speech fragments ($p < .01$, Bonferroni corrected for number of electrodes). On the sides are the individual cross-correlation profiles per electrode. The procedure for conducting the t-tests on the correlation values was identical to the previous analysis with correlation to speech spectral envelope (see Methods for more details).

## Filtering out background noise in speech fragments

Given the observed presence of speech tracking in dPCC, we addressed the question of how tracking of a continuous stream of speech is affected by additional sounds that often occur in natural situations. The question lends itself to being addressed since multiple scenes in the film contained dialogues with overlapping music, multiple people talking at the same time, sound effects, and distracting background noise ('noisy' fragments). We investigated how the responses in dPCC were affected by the mix of sounds in the speech stream. In particular, we assessed whether activity in dPCC evidenced responding to speech specifically or to the composite auditory input (speech plus other sources).

We obtained separate sound tracks of speech, music and sound-effects from the film producer (BosBros Productions, www.bosbros.nl). We selected a new set of speech fragments that contained speech combined with other sounds (n = 63). Of note, this set of 'noisy' speech fragments was only used in the present analysis and all other analyses were conducted with the previously selected 115 speech and 115 non-speech fragments. To investigate the effect of background noise, we tested whether dPCC responses to noisy speech fragments were more correlated to the speech envelope of the mixed track (what participants actually heard) or to that of the isolated speech track obtained from the film producer. The t-test comparing HFB correlation to the speech envelope in both tracks showed that some dPCC electrodes tracked the speech envelope of the isolated track significantly better ($\bar{\rho}_{S1} = .32 \pm .05$ and $\bar{\rho}_{S2} = .35 \pm .04$) compared to the mixed track ($\bar{\rho}_{S1} = .28 \pm .05$ and $\bar{\rho}_{S2} = .29 \pm .03$, **Figure 4.3**). The reported values show mean $\rho$-values and standard deviations over all significant electrodes ($\rho$-values were first averaged over all fragments separately for isolated or mixed track per electrode). None of the electrodes showed preference for the mixed track. This result suggests that dPCC was particularly sensitive to speech as opposed to music and noise.

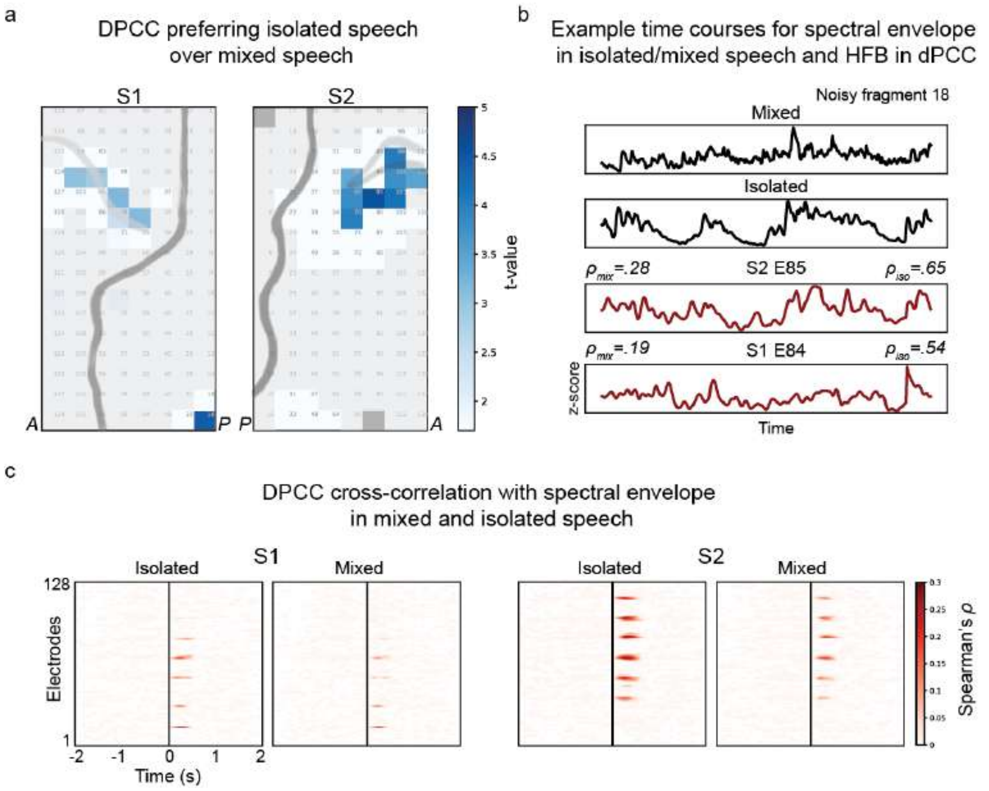


**Figure 4.3. Filtering out background noise from speech in dPCC. (a)** Results of the t-tests comparing the amount of correlation to speech spectral envelope in isolate speech or mixed sound track. The correlation and t-test procedures were identical to the ones previously described, except that the analyses were performed on a different set of fragments, which only included noisy speech and speech overlapping with other sounds (see Methods for more detail). **(b)** Example time courses for a noisy speech fragment. From top to bottom: spectral envelope of the fragment in mixed track, spectral envelope of the fragment in isolated speech track, HFB response of electrode 85 in S2 and HFB response of electrode 84 in S1. **(c)** Cross-correlation profiles averages over all noisy fragments separately for mixed and isolated speech tracks.

## Capturing the rhythmic phrasal structure of speech in dPCC

The dPCC region responded to speech isolated from the sound track in the speech condition and therefore must have been triggered by speech-specific properties. One of the auditory properties of speech that is particularly prominent in the absence of background noise is the rhythmic structure of speech. Specifically, when speaking, a continuous stream of speech is typically broken down into phrasal groups by the speaker. These groups are separated by pauses of at least 120-150 ms, but are highly variable in duration (Heldner, 2011; Zvonik and Cummins, 2003). Together, the switches between the groups of speech (phrases) and pauses create a rhythmic phrasal pattern, that constitutes one of the key perceptual characteristics of speech.

Following previous indications that dPCC could be involved in tracking of rhythmic properties of speech, including phrasal rates (Keitel et al., 2018), we tested whether in our study dPCC followed the rhythmic phrasal pattern in a continuous stream of speech. For this, we used the previously acquired manual linguistic annotation of the soundtrack (see Methods for details). The annotation contained onsets and offsets of every word in the sound. Using this information, we constructed a speech ON/OFF binary vector with ones corresponding to speech and zeros corresponding to pauses in a continuous stream of speech. Unlike the spectral envelope analysis, which compared tracking of the spectrotemporal structure in speech and non-speech fragments, here we focused on the binary structure of the speech envelope with phrasal groups (coded as 1) delineated by pauses (coded as 0). Thus, only the speech fragments were used (the previously selected 115 speech fragments). We fitted a linear regression to predict dPCC responses to speech fragments using the speech ON/OFF vector. The fit was significant for a large number of electrodes in both S1 and S2: max $F_{S1} = 2540$ and max $F_{S2} = 3330$ ($\mathrm{df}_1 = 45{,}998, \mathrm{df}_2 = 2$), $p < .001$, based on the permutation test (n = 10,000). Further inspection revealed a subset of dPCC electrodes with large positive $\beta$-weights indicating significant contribution of the speech ON/OFF vector to prediction of HFB responses in those electrodes: $\bar{t}_{S1} = 33.54 \pm 8.57$ and $\bar{t}_{S2} = 36.82 \pm 11.14$ (df = 45,998) at $p < .001$, Bonf. cor. (**Figure 4.4**). This result suggested that these dPCC electrodes followed the rhythmic phrasal pattern in a continuous stream of speech.

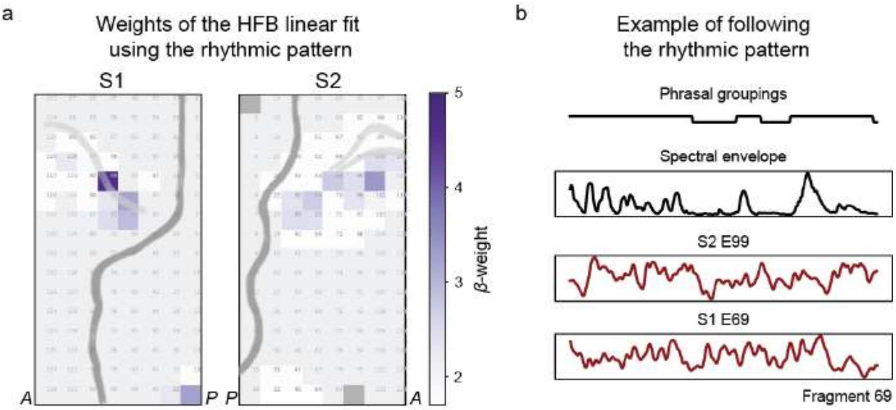


**Figure 4.4. Capturing the rhythmic phrasal pattern of speech in dPCC. (a)** Weights of a linear regression predicting HFB responses based on speech phrasal groupings (ON/OFF speech binary vector). Only speech fragments were used, thus OFF segments relate to pauses within a continuous stream of speech. **(b)** Example time courses for a speech fragment showing the rhythmic phrasal pattern in speech (ON/OFF speech binary vector), spectral envelope of the fragment and HFB responses in dPCC in S1 (E69) and S2 (E99).

## Capturing pitch of speech in dPCC

Another speech property of high perceptual relevance is pitch. Pitch is the perceptual correlate of the fundamental frequency of the speech signal. Pitch contour encodes changes in intonation of the phrase and voices of individual speakers are associated with distinct pitch ranges (Bishop and Keating, 2012; Collier, 1975). During speech, pitch is generated by vibrations of the vocal chords and is therefore a characteristic of only the voiced part of the speech signal. At the same time, being a frequency related characteristic of any auditory signal, pitch is not specific to speech and can be extracted from other signals such as music and environmental sounds, for example animal cries (Hevner, 1937; Tramo et al., 2005).

Since pitch and the spectral envelope are both related to the frequency component of speech, we first assessed the degree of interaction between them. We found that pitch and spectral envelope correlated significantly during speech fragments $(\bar{\rho} = .48, p < .001)$ and more than during non-speech fragments ($t = 7.95, \mathrm{df} = 228, p \ll .001$, **Figure 4.5a**). Moreover, the correlation was significantly higher for the isolated speech sound (speech-only track) compared to speech mixed with noise (mixed sound track, $t = 4.07, \mathrm{df} = 61, p \ll .001$). On the other hand, both pitch and spectral envelope inevitably also share information on the rhythmic structure of speech, such that for example during pauses both pitch and spectral envelope have near-zero values. To account for these interactions and isolate the effects of pitch and spectral envelope tracking, we examined the HFB residuals of the previous analysis (regression onto the rhythmic phrasal pattern) and computed their partial correlations with pitch (by taking spectral envelope into account) and spectral envelope (by taking pitch into account).

The partial correlation analysis showed that dPCC electrodes in both subjects tracked spectral envelope significantly better ($\bar{\rho}_{S1} = .14 \pm .01$ and $\bar{\rho}_{S2} = .14 \pm .01$) than pitch ($\bar{\rho}_{S1} = .01 \pm .02$ and $\bar{\rho}_{S1} = .02 \pm .02$) as assessed with paired t-tests per dPCC electrode: $t_{S1}$ranged from 4.11 to 10.23 and $t_{S2}$ ranged from 4.64 to 9.77 (df = 228) at $p < .01$, Bonf. cor. (**Figure 4.5b**). This result indicates that the activity of dPCC electrodes was more tightly related to the changes in the spectral envelope rather than pitch.

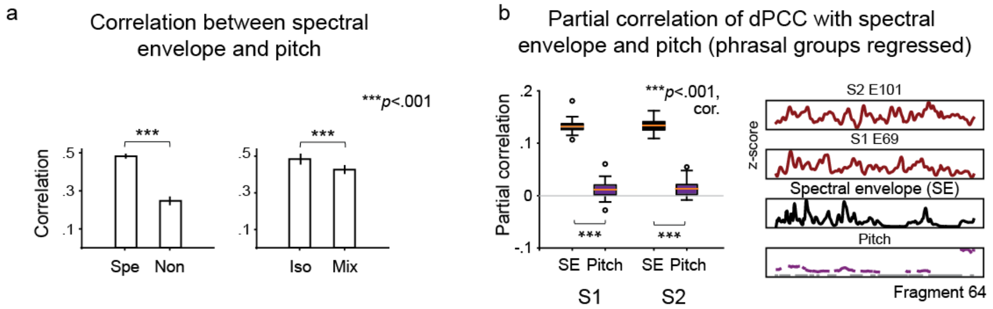


**Figure 4.5. Capturing pitch of speech in dPCC. (a)** Pearson correlation between speech spectral envelope and pitch. Left panel shows correlation during speech (*n=115*) and non-speech (*n=115*) fragments. Right panel shows correlation during noisy fragments (*n=63*) in isolated speech-only ('*iso*') and mixed ('*mix*') sound tracks. Error bars indicate the standard error of the mean. The reported difference between conditions (types of fragments on the left or sound tracks on the right) is significant at $\mathrm{p} < .001$. **(b)** Partial correlation of HFB responses with spectral envelope and pitch. Data from only the speech fragments were used. To account for the interaction with tracking of the rhythmic phrasal structure, the analysis was performed on the residuals of the linear fit of dPCC HFB responses to the binary rhythmic phrasal pattern (see results in b). Left panel shows the difference in partial correlation with spectral envelope and pitch for both subjects. Right panel shows example time courses of the electrodes with significant partial correlation to spectral envelope as well as the time course of the spectral envelope and pitch for speech fragment 64.

## Residual HFB responses to speech and non-speech fragments

Finally, having observed that dPCC activity reflects various properties of the speech signal, we asked whether the speech properties selected here were sufficient to explain the elevated dPCC response to speech compared to non-speech fragments. We also assessed if any of these perceptual auditory features could explain the responses of dPCC to non-speech sounds or whether their tracking was specific to the speech condition only.

For this, we computed a linear fit of the HFB responses in dPCC using the spectral envelope, rhythmic structure (audio sound being ON or OFF) and pitch information (*auditory properties*). The fit to the HFB responses was calculated separately for speech and non-speech fragments. From the fit we obtained the residual HFB responses separately for speech and non-speech conditions. These average residual responses were compared to each other and to the average HFB responses in the original neural data (prior to the linear fit on the auditory properties) using non-parametric alternatives to standard ANOVA tests (due to the likely violation of the requirement for the equal population variances, see Methods for details). First, we found an effect of speech condition regardless of whether we used the original HFB responses to speech or the residuals from the fit on the auditory properties: $F(1,76) = 167.57$ for S1 and $F(1,76) = 741.22$ for S2 (**Figure 4.6**). The difference in mean responses to speech compared to non-speech input became smaller for the residual HFB responses but it was still significant at $p < .001$. Second, we found that the effect of the fit to the auditory properties was only significant for the speech condition ($p = .001$, $m_{full} - m_{res} = .21$) and not for the non-speech condition ($p = .9$, $m_{full} - m_{res} \ll .001$). This indicates that auditory properties, such as spectral envelope, rhythmic structure and pitch could not explain HFB responses to non-speech input and therefore were not tracked during perception of non-speech sounds.

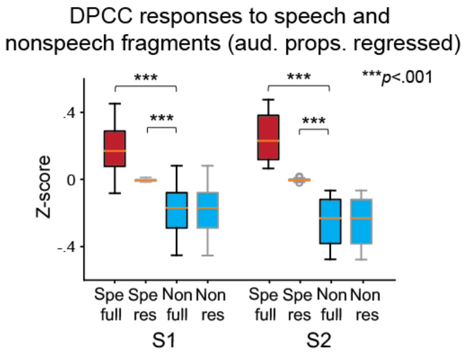


**Figure 4.6. Residual HFB responses to speech and non-speech fragments.** Comparison of the average HFB responses to speech (‘*spe*’, red boxplots) and non-speech (‘*non*’, blue boxplots) fragments. Original HFB data are used (‘*full*’, boxplots with black outline), as well as residual HFB data from the regression on all auditory properties (aud. props.): spectral envelope, pitch, rhythmic phrasal structure (‘*res*’, boxplots with grey outline). The comparisons between four types of data (‘*spe-full*’, ‘*spe-res*’, ‘*non-full*’ and ‘*non-res*’) were performed used non-parametric Kruskal-Wallis tests. The reported results are significant at $p < .001$.

## Functional specialization in PCC

The previous analyses established a clear connection between the activity within dPCC and speech perception of the feature film. Upon visual inspection, the location of the region appeared to overlap with part of the motor cortex associated with hand movement. To rule out the possibility that the observed effects could be explained by the visual perception of hand movement, we first assessed the interaction between hand movement and speech presence in the film. For this, we annotated all speech and non-speech fragments with hand presence and hand movement per frame (see Methods for details). Then, we used a $\chi^2$-test to assess the speech-hand interaction (**Table 4.1**). The test result was not significant, suggesting that there was no interaction between speech and hand conditions in the film data: $\chi^2(2,\ 609) = 1.31, p = .52$.

**Table 4.1. Contingency table for the $\chi^2$-test** assessing interaction of hand (in video) and speech (in audio) annotations.

| | Hand movement | Hand presence | No hand | Total |
|---|---|---|---|---|
| Speech | 114 | 61 | 107 | 282 |
| No speech | 127 | 62 | 138 | 327 |
| Total | 241 | 123 | 245 | 609 |

To determine a potential contribution to the brain signals by perceived hand movement in the film, a regression analysis was conducted with hand movement and brain signal in dPCC. The residuals were then entered in a t-test on HFB in speech and non-speech fragments, and dPCC tracking of the audio spectral envelope. Accounting for the hand movement did not significantly change the previously reported results as tested with two-sided Wilcoxon signed-rank tests ($Z_{S1} = -0.26, p = .79$; $Z_{S2} = -0.05, p = .96$ for the speech preference analysis and $Z_{S1} = -0.06, p = .95$; $Z_{S2} = -0.04, p = .97$ for the tracking of the spectral envelope analysis). All electrodes maintained significant after adding the covariate.

To assess how the location of speech-related activity compared with sensorimotor topography, data were used from hand movement and mouth articulation tasks performed by the same patients (see Methods for details). Results are displayed in **Figure 4.7a**, showing distinct functional specialization in PCC with posterior dPCC involved in hand movement, ventral PCC involved in mouth articulation and anterior dPCC involved in speech perception (**Figure 4.7b**). Inspection of cortical maps for individual fingers and speech articulators showed that the speech tracking electrodes overlapped most with the larynx articulation map (**Figures 4.1S** and **4.2S**). Of note, during the laryngeal motor task the subjects generated an audible humming sound and it is possible that the activity in dPCC could be related to tracking of the auditory feedback signal.

## Reproducibility of results with low-density ECoG grids

Finally, because HD grids are far less common in ECoG research than low-density grids (LDs, larger diameter and larger inter-electrode spacing), we sought to confirm some of our results with LD grids. Two participants with LD grids placed over PCC watched the same film (**Figure 4.8a**). We analyzed their HFB responses to the same speech and non-speech fragments and found a similar tendency for speech preference in anterior dPCC (**Figure 4.8b**). Notably, LD grid responses were associated with considerably lower t-values compared to HD grids: max $t_{S3} = 3.37$ and max $t_{S4} = 5.65$ versus max $t_{S1} = 27.85$ and max $t_{S2} = 27.74$. The analysis of speech tracking in LD grids (cross-correlation to speech envelope and STG electrodes) showed no significant results for dPCC (**Figure 4.8c**).

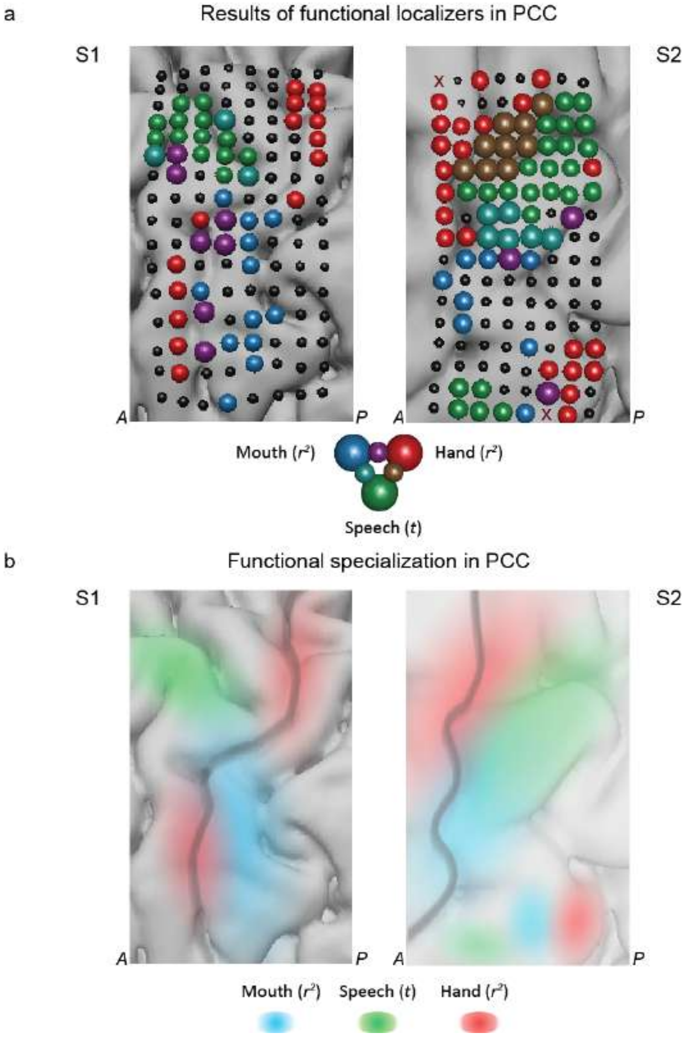


**Figure 4.7. Functional specialization in precentral cortex. (a)** Results of the functional localizers for mapping hand motor (red), mouth motor (blue) and speech perception (green) areas in sensorimotor cortex. 'Hand' localizer was a separate task with a block design with two conditions: 'flex fingers' and 'rest' ($r^2$ of HFB values). 'Mouth' localizer was a separate task with a block design with two conditions: 'move tongue inside mouth' and 'rest' ($r^2$ of HFB values). 'Speech' localizer is based on the results of the present study and shows comparison of two conditions: speech and non-speech fragments, also shown in Figure 1b (t-tests on HFB values). **(b)** A schematic representation of the functional specialization in precentral cortex. Dark grey line outlines the precentral sulcus. Shading refers to the function of the region: hand motor (red), mouth motor (blue) and speech perception (green).

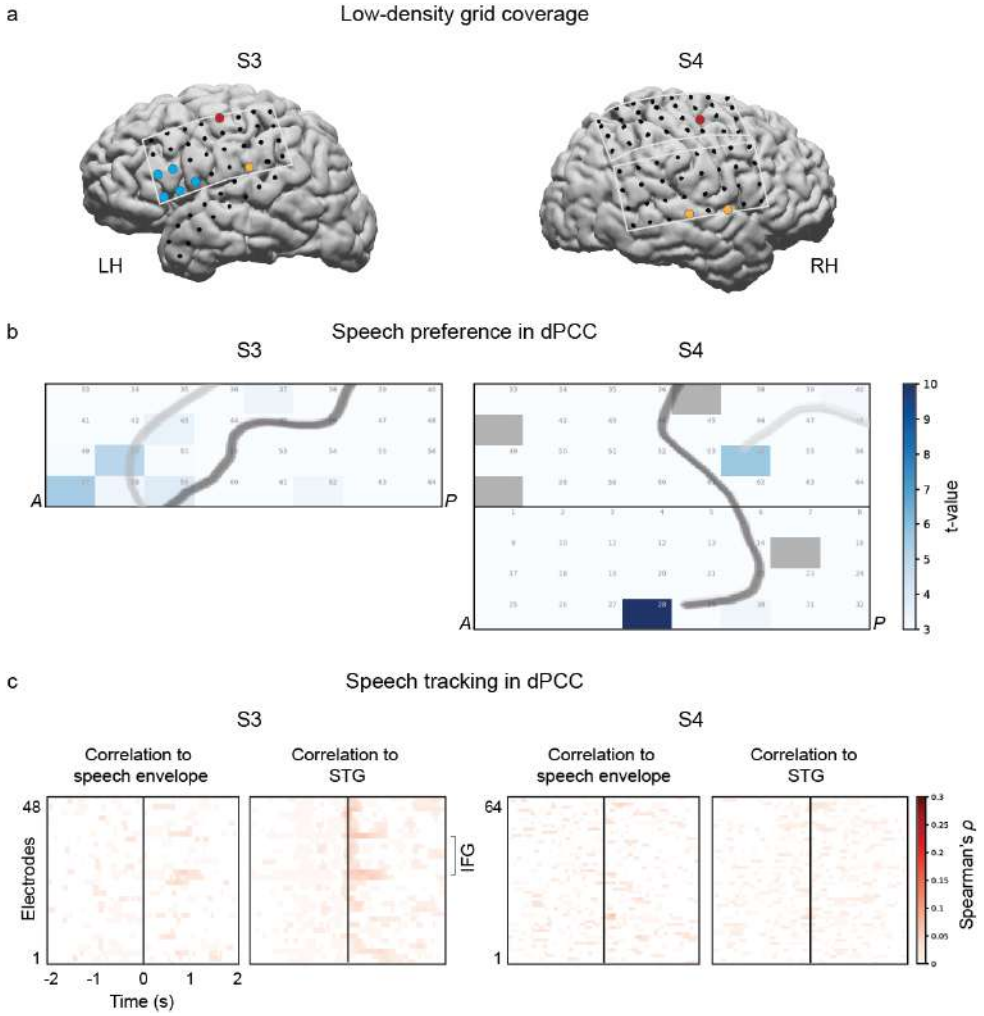


**Figure 4.8. Effects of the grid size. (a)** Brain coverage with low-density grids in two more participants (S3 and S4), who watched the same full-length feature film. Each dot represents an ECoG electrode. Some electrodes are colors based on their location/function: IFG (blue), STG (orange) and dPCC (red). Only the electrodes with significant t-values (comparing HFB in speech and non-speech fragments) are colored. **(b)** Results of the t-test comparing average HFB activity during speech and non-speech (noise, music) fragments (p < .05, Bonferroni corrected for number of electrodes). **(c)** Results of speech tracking. Shown are plots of cross-correlation between HFB and speech spectral envelope (left panel in both subjects) as well as plots of cross-correlation between HFB and a STG electrode (right panel in both subjects).

## 4.3 DISCUSSION

In the present study, we investigated and characterized neural responses in PCC to perceived natural speech, using HD intracranial recordings. We found that a region within dPCC (mostly its anterior portion) exhibited preference to perception of speech over other sounds. Groups of electrodes within this area displayed tracking of the speech spectral envelope, followed the speech phrasal patterns and filtered out music and noise. Combining these results with data from additional tasks, we were able to show that the cortical region outlined in the present results has a functional specialization distinct from hand motor and mouth articulation functions. Altogether, this work provides evidence that dPCC is actively involved in speech perception, although its role in perception is not yet clear. An additional finding was that the response characteristics were less clear in two patients with low-density grids, indicating that further research on speech perception (at least in this region) requires the use of HD intracranial electrodes.

### Defining dPCC involved in speech perception

The present findings provide strong evidence of the involvement of anterior dPCC in speech perception (**Figure 4.9a**). Previous research has implicated involvement of similarly located or neighboring regions during perception of speech with functional magnetic resonance imaging (fMRI; Du et al., 2014; Skipper et al., 2005), magnetoencephalography (Keitel et al., 2018), low-density ECoG (Cogan et al., 2014; Ding et al., 2016; Glanz et al., 2018) and transcranial magnetic stimulation (TMS; Floel et al., 2003; Meister et al., 2007). However, the exact location and therefore functional specificity of this region remain undefined. Various studies consider the region to be part of the premotor cortex rather than motor cortex (Glanz et al., 2018; Meister et al., 2007), even though some of the reported coordinates seem to belong to motor cortex proper according to the boundary delineated in a large meta-analysis study (Mayka et al., 2006). Some researchers point out that the observed effect is at the border of premotor and motor regions, corresponding to Brodmann areas 6 and 4 respectively (Wilson et al., 2004). Inconsistency extends to defining the boundaries of dorsal and ventral cortices as well (Mayka et al., 2006). Thus, several studies refer to the region as part of the dorsal (pre)motor cortex (Keitel et al., 2018; Meister et al., 2007), whereas others call it superior part of the ventral (pre)motor cortex (Cheung et al., 2016; de Heer et al., 2017; Glanz et al., 2018; Wilson et al., 2004). The boundary also differs between meta-analyses of neuroimaging and cytoarchitectonic data (Mayka et al., 2006; Rizzolatti and Luppino, 2001). The lack of distinct anatomical definition combined with a considerable variability in localization across individuals (Glanz et al., 2018) mark the challenge in delineating functional topography.

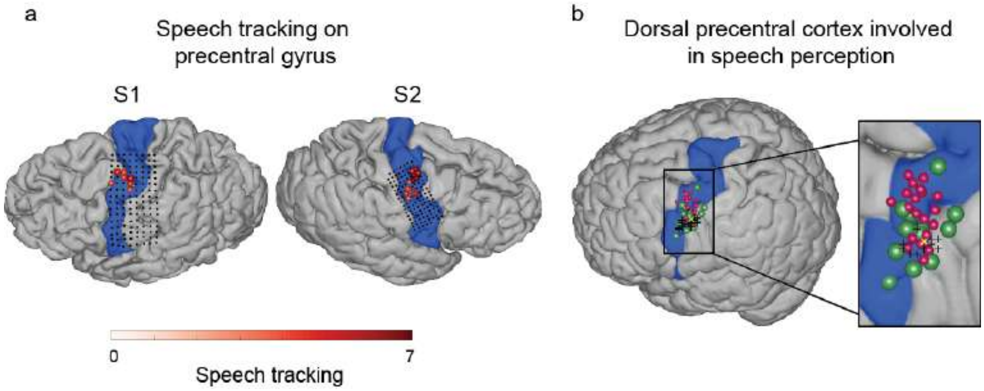

a
Speech tracking on precentral gyrus
S1
S2
0
7
Speech tracking
b
Dorsal precentral cortex involved in speech perception

**Figure 4.9. Defining dPCC involved in speech perception. (a)** Localization of the speech tracking results in the present study. The colored electrodes show significant tracking of the speech spectral envelope compared to the non-speech baseline (also shown of a grid plot in Figure 2a). The speech tracking colormap thus represents the t-values comparing the Fisher-transformed correlations to spectral envelope between speech and non-speech fragments (see Figure 2 and Methods for detail). The area corresponding to precentral gyrus is shown for each participant (highlighted in blue). Anatomical parcellation was performed in individual subject space using Freesurfer routines. **(b)** Comparison of the cortical localization of the present results to the literature (on the standard MNI brain). The present results from S2 (right hemisphere coverage) were projected onto the left hemisphere. The MNI coordinates of the present results are obtained using subject specific affine transformation matrices computed with SPM8 (MNI coordinates are reported in **Table 4.1S**). The results from the previous studies were projected on the standard brain surface using the MNI coordinates reported in those studies. Only left hemisphere coordinates were used. Area corresponding to precentral gyrus is shown (highlighted in blue). Anatomical parcellation was performed on the MNI brain using Freesurfer routines. Cortical tracking of the speech spectral envelope from (a) is shown (electrodes in red). Electrodes that show response to both perception and production of speech during real-world conversations from Glanz et al. (2018) are displayed in green. Center of the area responding to perceived syllables as measured with fMRI and reported in Wilson et al. (2004) is shown (yellow x). Finally, sites, whose stimulation with TMS affected perception of consonants reported in (Meister et al., 2007) are displayed as black crosses (one per individual subject).

Despite the difference in terminology, neural recording modalities and experimental paradigms, the present results show a considerable overlap in location with several previous reports (**Figure 4.9b**). Wilson et al. (2004) use fMRI to locate a site in dorsal precentral cortex that responds to perception of syllables. A recent study by Glanz et al. (2018) combined low-density ECoG and electrical stimulation mapping (ESM) to identify a region in superior ventral premotor cortex (adjacent to ours) that exhibited enhanced HFB responses during both production and perception of naturalistic speech. The present study differs from these two and the previously mentioned reports in several ways. First, we take advantage of HD neural recordings to obtain a detailed map of the function in dPCC. Second, we examined neural activity while participants were watching a feature film, eliminating constraints induced by a specific cognitive task. This is in contrast to Glanz et al. (2018) who also used a naturalistic experimental setup but analyzed speech perception and production moments in real-world conversations. Thus, our results cannot be attributed to motor planning or predictions as part of a face-to-face interaction and are due to speech perception proper. This approach made it possible to associate a distinct portion of dPCC with a processing of features that are specific to speech. The specificity revealed by the tracking of the spectral envelope and phrasal structure of speech has, to the best of our knowledge, not been reported before.

### Relation to hand and mouth motor processes in dPCC

The present results show reliable activation of anterior dPCC by perceived speech. However, given previous research and the region's location in the brain, it is important to consider our findings in the context of motor processing. In particular, the mirror neuron theory implicates neurons in (pre)motor cortex in both perception and execution of goal-oriented action, particularly emphasizing their role in action understanding (Di Pellegrino et al., 1992; Rizzolatti and Craighero, 2004). Even though the mirror neuron theory has met considerable criticism (Hickok, 2009), many researchers continue to agree that the observed neural activity of (pre)motor region in both human and non-human primates reflects some form of interpretation of the perceived actions (Salo et al., 2019).

The dPCC is primarily associated with hand motor processing, and one could argue that the present results could be attributed to the fact that the motor cortex merely responds to perceived communicative hand gestures. We find this explanation unlikely for several reasons. First, the present results rely on correlations of the dPCC HFB activity to the speech spectral envelope, which captures the slowly varying shape of the speech signal. The spectral envelope has been shown to be critical for perception of individual phonemes as well as overall sentence comprehension (Arai et al., 1996; Ter Keurs et al., 1993). Many core regions involved in speech processing show tracking of this speech feature (Kubanek et al., 2013). Second, accounting for the hand presence and movement in the film frames did not change the present results in any of the electrodes, suggesting that anterior dPCC is unlikely to respond to perceived actions and hand movements, but is rather related to tracking of the speech-specific information. Finally, utilizing the high spatial resolution of HD electrodes we were able to map individual hand movements on the dPCC cortex, and found their location to be different from the area that tracked perceived speech (**Figures 4.7** and **4.2S**).

The notion of the motor cortex supporting both action perception and execution is at the core of the motor theory of speech perception. It posits that the cortical regions implicated in mouth articulation (ventral PCC, 'face area') and motor planning (ventral premotor cortex) could subserve simulation and phonological prediction processes during speech perception (Cheung et al., 2016; Pulvermüller et al., 2006; Skipper et al., 2007). However, the region found in our study is located considerably more superior to the mouth motor region in both subjects, indicating that it is separate from ventral PCC proper (**Figures 4.7** and **4.1S**). At the same time, there appeared to be a considerable overlap with the dorsal laryngeal motor region identified in this study (**Figure 4.1S**), which has recently been reported to be related to volitional control of pitch (Dichter et al., 2018). Here, we show that the identified region tracked properties in perceived speech beyond pitch (**Figure 4.5b**). Moreover, regressing various acoustic features from the neural responses (including pitch) did not account entirely for the dPCC elevated response to speech compared to non-speech sounds (**Figure 4.6**). Altogether, this evidence suggests that either the identified dPCC region has other function in addition to laryngeal motor control or the currently considered laryngeal function of dPCC should be revised. Of note, it is currently considered that there are two laryngeal motor regions: one in dPCC and another one in ventral PCC (Bouchard et al., 2013; Brown et al., 2007; Simonyan and Horwitz, 2011). Only the dorsal region tracks perceived speech in our study.

### The role of dPCC in speech perception

The finding of a distinct region, just anterior to the ‘hand knob’ and superior to the ‘face area’, that tracks the speech spectral envelope raises questions about its function. Although it appears to overlap with the laryngeal motor area (Dichter et al., 2018), it is far from obvious why the larynx would specifically be involved in speech perception whereas other mouth articulators are not (see **Figure 4.1S**). Several ideas about the function of this region have been previously reported, including generation of forward motor representations of speech sound, facilitatory mechanisms for perception under difficult conditions and a role in prediction and processing of temporal information in speech.

Meister et al. (2007) reported that repetitive TMS stimulation of the premotor region (dorsal and anterior to the central sulcus) leads to a significant decline in subjects’ ability to discriminate consonant sounds presented in noisy conditions. The authors suggested that this area is crucial for mapping of acoustic representations of speech sounds onto corresponding articulatory gestures. They theorized that premotor cortex might feed these top-down motor representations forward to STG for their comparison against the acoustic input and thus have a causal role in speech perception. In the present study, we find consistent involvement of the same region (dorsal and anterior to the central sulcus) in tracking of perceived continuous speech. At the same time, we do not find support for the notion of feedforward processing from dPCC to STG given that we do not observe any consistent lag between HFB activity in dPCC and STG that one would expect in a feed-forward theory.

An alternative theory explaining activation of (pre)motor cortex during speech perception is based on the hypothesis of its facilitatory rather than causal function during perception of speech. This theory is based on activation of (pre)motor cortex during perception of noisy and degraded speech (Callan et al., 2004, 2003; Du et al., 2014). The facilitation effect is thought to be achieved through sensorimotor integration and engagement of an internal model that maps speech sounds to articulation. Although the effect was reported to be localized to the ventral premotor cortex (Du et al. (2014), the speech perception maps for both noisy and clean speech seem to include dPCC as well. Sato et al. (2009) suggested that the facilitatory function of (pre)motor region should manifest in one of two scenarios: (1) when varying task complexity (Du et al., 2014; Wilson and Iacoboni, 2006) or (2) during conversational exchange (Foti and Roberts, 2016; Scott et al., 2009). Neither option can fully account for the activation of this region during passive perception of a narrative, such as a feature film. It is possible that the multisensory integration during the audiovisual perception can contribute to the present results. However, this would mean that various reports of dPCC involvement are due to different conditions: noise and task difficulty (Meister 2007, Sato 2009), conversation exchange (Glanz et al., 2018), and passive listening to naturalistic speech (present work). We tend to consider this unlikely and instead believe that dPCC simply elicits a more basic, fundamental response to perceived speech.

A different line of work associates the activation in dPCC with cortical entrainment to rhythmic features of individual sentences (Bengtsson et al., 2009; Ding et al., 2016; Keitel et al., 2018). In this view, dPCC involvement in speech perception ultimately connects to temporal prediction and tracking of rhythmic structure in (pre)motor cortex (Chen et al., 2006; Morillon et al., 2014). We find that anterior dPCC follows the phrasal grouping patterns in the continuous stream of speech. At the same time, regressing out the temporal acoustic properties from HFB responses does not entirely remove the difference in activity associated with perception of speech compared to non-speech sounds (**Figure 4.6**). Moreover, there appears to be a clear encoding of the frequency components of speech (spectral envelope and pitch). This indicates capturing of information beyond the rhythmic structure in dPCC and aligns with results from Meister et al. (2007) where the inhibition of dPCC with repetitive TSM affects the subjects' ability to discriminate between perceived consonants – a task that is free from any temporal pattern.

In sum, findings in literature as of yet do not provide a comprehensive account for the function of dorsal (pre)motor/precentral cortex in perception of speech. Apart from theories embedding the function of dPCC in some form of motor processing, other reports assign dPCC responses to speech in terms of semantic (de Heer et al., 2017), lexical (Duffau et al., 2003) and verbal memory (Müsch et al., 2018) processing. A more unifying theory of the function of this region that could explain its involvement in both perception and production of speech (Glanz et al., 2018; Wilson et al., 2004) remains much needed in the field. We believe our current findings provide some direction to this endeavor.

With the spatial and temporal resolution of the recordings in the current study, a close association between a distinct dPCC region and processing of perceived speech was observed. To better understand the functional relevance of this, further investigation of functional sequelae following virtual lesioning in this area seems warranted, for instance with ESM over HD grid electrodes. Less specific virtual lesion techniques such as TMS have elucidated some of the relationship between speech perception, speech production and hand motor cortices (see a review Möttönen and Watkins, 2012). An overlap in speech production and perception was observed by stimulating through standard low-density electrode grids in dPCC (Glanz et al. (2018). This study, however, also reported motor effects spanning multiple body parts including lips, tongue, neck, eyes, chin, head and fingers, suggesting perhaps a lack of specificity of stimulation. Only the study by Dichter et al. (2018) reports on ESM over HD grid electrodes, where a topographically similar region displayed motor effects on the larynx (other articulators were not tested). A study on patients with lesions in premotor cortex showed that electrical stimulation of dPCC lead to no articulation deficits but rather difficulties in object naming (Duffau et al., 2003). Given that the spatial resolution of TMS and standard 10 mm density ECoG is in the range of 1 cm (Roth and Hallett, 1992; Thielscher and Kammer, 2002), it may well be that separation and in-depth investigation of functions in the regions we report on, requires more specific stimulation (HD grid). Given our findings, a focus on separating motor and language functionality would be of interest.

## Limitations

The present work has a number of limitations. For one, due to the rarity of HD recordings data from only two HD participants were available. Complimentary work with low-density grids suggests that HD recordings are necessary for accurate mapping of function in dPCC.

Second, it is possible that the present results were confounded by perception of the visual stream of the film and particularly by perception of movement. By labeling hand presence and movements in the film, we could correct for the latter confound to some degree. Analyses of the data after removing any hand presence or movement did not affect the statistical results, suggesting that the confound was minimal at best.

Related to this, it is possible that eye movements could have interfered in comparing brain activity during speech and non-speech fragments. However, it is unlikely for eye movements to correlate with various auditory properties of speech including frequency-based characteristics such as spectral envelope and pitch.

## Conclusions

We investigated the involvement of precentral cortex (PCC) in continuous speech perception using high-density (HD) intracranial recordings. Our results show that a specific region within the dorsal portion of PCC (dPCC) tracks various properties of speech including, but not limited to, its spectral envelope, pitch, and rhythmic phrasal groupings even with additional background noise or sounds. Tracking occurs in parallel to the activity in superior temporal cortex. The location of identified region is distinct from the hand motor and mouth articulator areas. In addition, we find that these results are more pronounced in HD grid participants compared to standard intracranial grids, indicating the importance of both spatial and temporal detail in studying neural responses to speech perception on sensorimotor cortex.

## 4.4 MATERIALS AND METHODS

### Participants

All participants were admitted for diagnostic procedures with medication-resistant epilepsy. They underwent subdural electrode implantation (low-density clinical grids) to determine the source of seizures and test the possibility of surgical removal of the corresponding brain tissue. In two subjects, additional high-density (HD) grids were placed over sensorimotor cortex for research, after approving the procedure and signing consent. Research could be conducted between clinical procedures. All patients gave written informed consent to participate in accompanying electrocorticography (ECoG) recordings and gave permission to use their data for scientific research. The study was approved by the Medical Ethical Committee of the Utrecht University Medical Center in accordance with the Declaration of Helsinki (2013).

### Film stimulus

A Dutch feature film "Minoes" (2001, BosBros Productions, www.bosbros.nl) was used as a stimulus for the film-watching experiment. The film was 93 minutes long (78 minutes before credits) and told a story about a cat Minoes, who one day transforms into a woman. In her human form, she meets a journalist Tibbe. Together, they solve several mysteries involving their town and during their adventures eventually fall in love. The film was made in Dutch and was easy to follow for all ages. Patients reported that they had enjoyed watching the film.

### ECoG experiment

Four patients (age 27±8, 3 females) watched the film. Two patients were implanted with grids in the left hemisphere, and two in the right. All patients had left hemisphere as language dominant, based on fMRI or the Wada test (**Table 4.2**).

Two patients were implanted with high-density (HD) grids over the sensorimotor region: S1 (128 contacts, 1.2 mm exposed diameter, inter-electrode distance 4 mm, left sensorimotor cortex) and S2 (128 contacts, 1 mm exposed diameter, inter-electrode distance 3 mm, right sensorimotor cortex). The suspected pathological regions in these patients did not extend to the sensorimotor region covered by the HD grids. This was clinically confirmed after implantation. Two remaining patients (S3 and S4) were only implanted with LD clinical electrode grids (2.3 mm exposed diameter, inter-electrode distance 10 mm, between 48 and 128 contact points). LD grids had perisylvian coverage including frontal and motor cortices. Patient-specific information about the grid hemisphere, number of electrodes, and cortices covered is summarized in **Table 4.2**.

**Table 4.2. Electrode grid information for all participants (both HD and clinical low density).** Shown is information about the number of electrodes, grid hemisphere, covered cortices, handedness, and language-dominant hemisphere per patient. L, Left; R, right; F, frontal cortex; M, motor cortex; T, temporal cortex; P, parietal cortex; O, occipital cortex; fMRI, functional magnetic resonance imaging.

| Patient | N of electrodes | Grid hemi-sphere | Cortices covered | Handed-ness | Language dominance | Grid |
|---|---|---|---|---|---|---|
| S1 | 128 | L | F, M, T | R | L (fMRI) | HD |
| S2 | 128 | R | T, P, O | R | L (Wada) | HD, LD |
| S3 | 64 | R | F, M, T, P | R | L (fMRI) | LD |
| S4 | 64 | L | F, M, T, P | R | L (fMRI) | LD |

In the experiment, each patient was asked to attend to the film displayed on a computer screen (21 inches in diagonal, at about 1 meter distance). The stereo sound was delivered through speakers with the volume level adjusted for comfort for each patient. Due to the long duration of the film, the patients were given an option to pause the film and quit the experiment at any time. In that case, the patient could continue watching the film at a later time starting from the frame they had paused on.

During the experiment, LD ECoG data were acquired with a 128 channel recording system (Micromed) at a sampling rate of 512 Hz filtered at 0.15–134.4 Hz. HD ECoG data were acquired with a separate system (Blackrock, Blackrock Microsystems) at a sampling rate of 2,000 Hz filtered at 0.3–500 Hz. The film was shown using Presentation software (Neurobehavioral Systems), which allowed us to synchronize the film sound with the ECoG recordings. In addition, audio-visual recordings of the room, patient and computer screen were collected and used to confirm synchronization.

## ECoG data processing

All electrodes with noisy or flat signal (based on visual inspection) were excluded from further analyses. After applying a notch filter for line noise (50 and 100 Hz), common average rereferencing was applied per patient, separately for LD and HD grids. Data were transformed to the frequency domain using Gabor wavelet decomposition at 1–125 Hz in 1 Hz bins with decreasing window length (four wavelength full-width at half maximum). Finally, high frequency band (HFB) amplitude was obtained by averaging amplitudes for the 65–125 Hz bins and the resulting time series per electrode were downsampled to 100 Hz. Electrode locations were coregistered to the anatomical MRI in native space using computer tomography scans (Branco et al., 2018; Hermes et al., 2010) and FreeSurfer (Fischl, 2012). The Desikan–Killiany atlas (Desikan et al., 2006) was used for anatomical labeling of electrodes in LD grids (closest cortical structure in a radius of 5 mm).

## Linguistic annotation

The soundtrack of the film was extracted using Audacity software (Audacity Team). The stereo track was merged into a mono track and downsampled to 16 kHz. This audio track was used for linguistic annotation.

From the film production company we obtained film subtitles and film script. These were used to produce a preliminary text-to-audio alignment in Praat (Boersma and Weenink, 2016). The alignment was created automatically by converting subtitle text into Praat annotations based on the subtitle time stamps. The subtitle text was compared against the script and corrected accordingly. Then, a number of undergraduate students were employed to correct the automatic text-to-audio alignment. Each student corrected the time markers of the subtitle text and created a tier with markers for onsets and offsets of individual words. Additionally, the students marked moments of overlap between speech and other sounds, such as music and audible noise. The students received detailed instructions regarding the waveform and spectrum properties of sound that could aid in determining the onsets and onsets of individual words. A trained linguist further verified their manual annotation. As a result, we obtained a linguistic annotation file with three tiers: subtitle text (1), individual word boundaries (2), overlap between speech and music or noise (3).

In addition, in the moments of the film with no speech present we annotated moments of presence of other sounds: music and various noises. These contained no overlap with speech and were used for extraction of non-speech fragments.

## Audio processing

The sound spectral envelope was extracted from the film soundtrack using NSL toolbox (Chi et al., 2005). We first extracted a sound spectrogram following the biological model of sound processing by the cochlea (Chi et al., 2005). The spectrogram was extracted at 8 ms frames along 128 logarithmically spaced frequency bins in the range of 180 – 7200 Hz. We then averaged the spectrogram data over the frequency bins to obtain a 1D spectral sound envelope. The resulting spectral envelope was downsampled to 100 Hz to match the sampling rate of the ECoG HFB time courses.

In addition, for pitch related analyses we extracted pitch contour from the film soundtrack using an autocorrelation algorithm (Boersma, 1993) as implemented in Praat. We used the default parameters for pitch estimation.

In subsequent analyses we assessed the difference in neural processing of speech and non-speech sounds. For this, we extracted a set of speech and non-speech fragments of the sound track based on the manual linguistic annotation. Each fragment was a continuous 4-second long fragment of the soundtrack. In total, we extracted 115 speech and 115 non-speech (containing music, noises, etc.) fragments. In case of speech fragments, we allowed pauses between speech instances within a fragment of no longer than 500 ms.

Additionally, for further analyses on tracking of speech in noisy conditions we compared the amount of speech tracking in HFB responses in mixed sound track (what patients actually heard) and isolated speech track (speech-only track obtained directly from the film company). The isolated speech track was processed the same way as the mixed sound track (extracted from the film as described above). Thus, we obtained the sound envelope for the isolated speech track from the sound spectrogram and downsampled it to 100 Hz. In addition, for these analyses we selected a set of noisy speech fragments (i.e. with audible overlap of speech with music and noise sounds, n = 63), based on the manual linguistic annotation. There was some overlap (17 fragments) between these 63 noisy fragments and the set of previously defined 115 speech fragments.

### Preference to speech fragments in dPCC

Independent two sample t-tests were used to compare average HFB amplitude values during speech and non-speech fragments (n = 115). For this, per electrode we averaged HFB responses over the entire fragment and compared the vector of 115 HFB values in speech fragments against the vector of 115 HFB values in non-speech fragments. The t-tests were conducted individually per each electrode. The p-values were computed parametrically and were corrected for multiple comparisons using Bonferroni correction for the number of electrodes per subject.

Because of the large number of electrodes per subject (n = 128 in both S1 and S2), the outcome of this analysis was used to limit the amount of comparisons in further analyses. Thus, all further analyses involving HFB responses to the film were performed only in the subset of electrodes that showed preference to speech fragments (electrodes with significant t-values from this analysis). This limited the number of multiple comparisons to 20 electrodes in S1 and 41 electrodes in S2.

All statistical testing for this and further analyses was conducted using *numpy* (Oliphant, 2006), *scipy* (Jones et al., 2001), *scikit-learn* (Pedregosa et al., 2011) and *statsmodels* (Seabold and Perktold, 2010) libraries for Python.

## Tracking of speech spectral envelope in dPCC

### Correlation to spectral envelope of speech

First, both sound spectral envelope and HFB data were rank-transformed within selected fragments: all original values were assigned a rank from 1 to *k* within a fragment based on the value magnitude. All further computations were performed on the ranked data rather than the original values as the rank transformation preserved the monotonic rather than linear relationships in the data. A Pearson correlation coefficient was computed on the rank-transformed sound envelope and HFB data at all lags (in steps of 100 ms). Since the correlation was calculated on ranked data, the resulting score corresponded to the non-parametric Spearman correlation coefficient ($\rho_e$):

$$\rho_e = \frac{\mathrm{cov}(\mathrm{rx_e},\ \mathrm{ry})}{\sigma_{\mathrm{rx_e}}\sigma_{\mathrm{ry}}} \tag{4.1}$$

where $\mathrm{rx_e}$ and ry are rank-transformations of $\mathrm{x_e}$ (HFB response per electrode) and y (audio envelope), respectively, and $\sigma_{\mathrm{rx_e}}$ and $\sigma_{\mathrm{ry}}$ are the standard deviations of the rank variables.

The maximal correlation was determined per fragment from all the biologically plausible lags in the range of -100 to 500 ms around the sound onset. The maximal correlation scores were Fisher-transformed ($z_e$) prior to further comparisons:

$$z_e = \frac{1}{2}\ln(\frac{1+\rho_e}{1-\rho_e}) \tag{4.2}$$

Independent two sample t-tests were used to assess the statistical difference between the average HFB correlation to sound spectral envelope in speech and non-speech fragments. The statistical significance was assessed parametrically and the p-values were corrected for the number of electrodes in the analysis (20 electrodes in S1 and 41 electrode in S2).

### Correlation to STG electrodes

First, per subject we identified a STG electrode (from the same HD grid in S1 and from a LD grid in S2) with the highest Spearman correlation to the audio envelope. This correlation procedure was identical to the one described above. Then, the time course of the selected STG electrode was cross-correlated to the dPCC electrodes (also through the similar correlation procedure, except that the maximal STG-dPCC correlation was taken within the range of -200 to 200 ms). This range was chosen because the previously reported lags of brain response to speech perception was in the range of 200 – 400 ms for both PCC and STG (Cheung et al., 2016; Glanz et al., 2018; Kubanek et al., 2013). Thus the lag between the two regions should be within the chosen range. Independent two sample t-tests comparing correlation in speech and non-speech fragments as well as the statistical significance were performed in the similar fashion as described above.

### Filtering out background noise in speech fragments

Only the previously extracted noisy fragments (n = 63) were used (see Audio processing section). Spearman cross-correlation and paired sample t-tests were used to compare HFB correlation to the speech envelope in isolated and mixed sound tracks. The procedure followed the previously described Spearman correlation/t-test pipeline. However, instead of comparing correlations during speech and non-speech fragments, we compared correlation to speech envelope in isolated speech and mixed sound tracks and thus used paired sample t-tests. Apart from that, all procedures were identical to the correlation and t-test procedures described above. The range of -100 to 500 ms was used to identify the maximal correlation to the sound envelope.

### Capturing the rhythmic phrasal pattern of speech in dPCC

To determine whether HFB responses in dPCC followed the phrasal grouping patterns in speech we first constructed a binary speech ON/OFF vector. All previously used speech fragments (n = 115) were concatenated. Using the manual linguistic annotation we assigned a value of 1 to all time points during speech and a value of 0 to all time points where speech was absent (= pauses in a continuous stream of speech). Then, a linear regression was used to predict the z-scored HFB responses ($\mathrm{y_e}$) using this binary speech ON/OFF vector (x):

$$\mathrm{y_e} = \beta_e{}^{\mathrm{T}}\mathrm{x} + \varepsilon_\mathrm{e} \tag{4.3}$$

where $\varepsilon_\mathrm{e} \sim \mathcal{N}(0, \sigma)$.

Since we concatenated all speech fragments we fit a single regression model for all data rather than fitting an individual regression model per speech fragment. The statistical significance of the fit was assessed using F-tests (with the null hypothesis that all regression $\beta$-weights were equal zero) and permutation testing for determining the chance threshold of the F-statistic. During the permutation testing we permuted the order of the speech fragments prior to their concatenation 10,000 times and each time fit a new linear regression on the permuted speech ON/OFF vector. Then we compared the F-statistic of the actual fit to the 99.999th percentile of the permutation distribution, which corresponds to a chance level of .001. The significance testing procedure was repeated per electrode.

## Capturing pitch of speech in dPCC

Given that both pitch and spectral envelope reflect the spectrotemporal properties of speech, prior to the analyses on the neural data we assessed the amount of shared information between the two auditory features. For this, we computed correlations between pitch and spectral envelope in speech and non-speech fragments separately. We calculated both Pearson and Spearman correlation coefficients, but the results were comparable between the two. The values displayed in **Figure 4.5a** represent the Pearson correlation values. The amount of correlation was significant in both speech and non-speech conditions (as tested with one-sample t-tests on the Fisher-transformed correlations). The difference in correlation between speech and non-speech fragments was assessed using an independent two-sample t-test on the Fisher-transformed data. In addition, we also computed the amount of correlation between spectral envelope and pitch for the speech noisy fragments, using either isolated speech-only sound track or the mixed sound track. The difference in correlation between the two tracks was also assessed with a paired two-sample t-test on the Fisher-transformed correlation data.

For the analyses on the neural data, we aimed to account for the interactions between the spectral envelope, rhythmic phrasal structure and pitch. For this, we used the residuals of the previous analysis fitting the binary ON/OFF speech vector (rhythmic phrasal structure) to the HFB responses and computed partial correlations of the HFB residuals with pitch and spectral envelope. Thus, per ECoG electrode (same selection of electrodes as in all previous analyses), we computed a partial Spearman correlation with pitch while accounting for the spectral envelope data (1) and with spectral envelope data while accounting for pitch (2). The analysis was only performed on the data from the speech fragments.

## Analysis of residual HFB responses to speech and non-speech fragments

Finally, we assessed the difference in the HFB responses during speech and non-speech fragments by taking into account the gained knowledge about HFB tracking of the auditory properties of the input audio signal. We fitted an ordinary least squares solution to predict the HFB responses based on all previously used auditory properties (spectral envelope, rhythmic phrasal structure, pitch). The fit was computed separately for speech and non-speech fragments. In non-speech fragments, the 'rhythmic' binary vectors also captured pauses and similar to the speech condition represented the sound being ON or OFF. Pitch and spectral envelope were calculated the same way as for the speech fragments.

The HFB residuals of the fit using auditory properties were compared between speech and non-speech condition. For completeness, we also included the comparisons with the original HFB responses to speech and non-speech fragments (HFB data prior to the fit, same data as used in the first t-test analysis comparing average responses to speech and non-speech fragments). HFB data were z-scored prior to the fit. Because we aimed to compare the average HFB amplitude between the original data ('full') and same data after regressing the auditory properties ('residuals') in both conditions (speech and non-speech) we refrained from using parametric approaches such as a one-way ANOVA test. It seemed logical to assume that in case of a successful fit, the original data and the residuals would not have equal population variances. Instead, we opted for a non-parametric test based on ranked transformations of the data, such as a Kruskal-Wallis test. In case of the rejection of the null hypothesis that all groups had equal means, we performed post-hoc tests determining which groups of data showed significant difference in means while accounting for the multiple comparisons (non-parametric post-hoc Dunn's tests). In addition, since the groups were clearly organized along two factors (type of fragments: speech and non-speech, and type of used data: full or residual HFB responses), we aimed to investigate the main effects of each factor and their interaction. To account for the violation of the assumption about equal population variances and to follow the logic of the Kruskal-Wallis and Dunn's tests, we performed a two-way factorial ANOVA analysis (which was simply equivalent to a linear regression using categorical factor variables) on the rank-transformed neural data.

## Functional specialization in PCC

### Testing interference from visual hand perception

All the speech (n = 115) and non-speech (n = 115) fragments used in the analyses were annotated with respect to hand presence and movement. For annotation we used ELAN software (Brugman et al., 2004), which unlike Praat supports a video stream. We went through every frame corresponding to speech and non-speech fragments and annotated it with hand movement using a three-level scale: 0 – no hand presence, 1 – hands are visible but there is no movement, 2 – clear hand movement. A $\chi^2$ analysis was employed to test interaction between speech and hand variables across the used fragments. The main $\chi^2$ test reported in the Results assessed interaction between two levels of speech variable ('speech present' and 'speech absent') and three levels of hand variable ('hand moving', 'hand present' and 'hand absent'). In addition, we performed another $\chi^2$ analysis with a simplified hand variable that contained only two levels ('hand present' and 'hand absent') by replacing all 'hand moving' annotations with 'hand present' annotations. There was no interaction between hand and speech variables as a result of this analysis either: $\chi^2(2,\ 609) = 0.97, p = \ .32$.

In addition, we also used the hand movement annotation (the three-level scale one) as a covariate in two previous analyses: the speech preference analysis (t-test on average HFB amplitude in speech versus non-speech fragments) and the tracking of the spectral envelope analysis (cross-correlation of HFB to spectral envelope). For this, we constructed a vector of hand movement/presence/absence values per each fragment (both speech and music fragments were used, thus 230 fragments in total). The data were concatenated across all fragments and the least ordinary squares solution was applied to fit HFB data using the hand regressor values. The obtained residuals of the linear fit were used to repeat both the speech preference and the tracking of the auditory envelope analyses. For both analyses new t-statistics (comparing speech and non-speech conditions) were obtained using the residual HFB data. These updated t-statistics were compared against the original t-statistics (obtained from the HFB data without regressing out the hand annotation). The statistical comparisons of the original and updated t-statistics were performed using non-parametric two-sided Wilcoxon signed-rank tests. Only electrodes with significant original t-statistics were used in these comparisons.

### Relation to the hand motor and mouth articulator localizers

Both HD patients performed localizer tasks to identify cortex involved in hand motor and mouth motor execution. The hand motor task has a finger movement task with a randomized event-related design. The task was previously used with fMRI and ECoG to obtain cortical representation of finger movement (Siero et al., 2014). Each patient was instructed to flex the thumb, index or little finger of their right hand depending on the cue. Each trial consisted of two flexions of one finger. During 'rest' trials the patients were instructed to remain still. The data from both subjects were preprocessed (bad channel rejection, line noise removal) and responses in HFB (65 – 125 Hz) were extracted. The data from three finger movement conditions ('thumb', 'index' and 'little finder') were all treated as the single 'move' condition. The 'move' condition trials were compared against the 'rest' trials using a signed $r^2$ measure (**Figure 4.7a**). The reported $r^2$ values were significant at $p \ll .001$ in each subject.

The mouth articulator task also had a randomized event-related design. The task was previously used with fMRI and ECoG participants to identify cortical sites involved in articulation (Bleichner et al., 2015). Each patient was instructed to move different parts of their mouth involved in articulation: lips, tongue, jaw or larynx depending on the cue ('lip', 'tongue', 'teeth clench' and '*mmmh*' respectively). During 'rest' trials the patients were instructed to remain still. The data from both subjects were preprocessed (bad channel rejection, line noise removal) and responses in HFB (65 – 125 Hz) were extracted. The data from four articulator movement conditions (lips, tongue, jaw and larynx) were all treated as the single 'move' condition. The 'move' condition trials were compared against the 'rest' trials using a signed $r^2$ measure (**Figure 4.7a**). The reported $r^2$ values were significant at $p \ll .001$ in each subject.

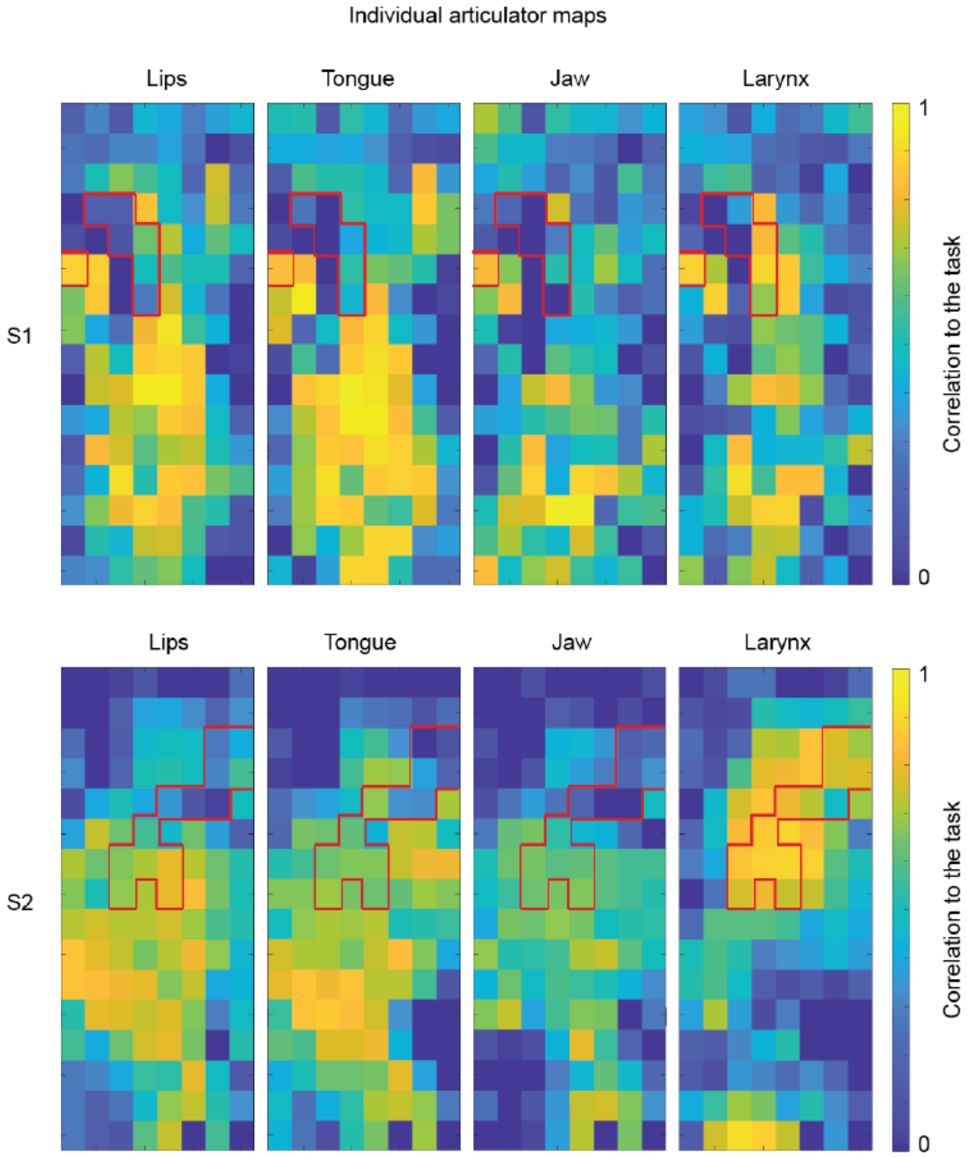
### Reproducibility of results with LD grids

In addition to the main analyses on HD recordings, data from two patients who were only implanted with LD clinical grids were analyzed. Similar to the analyses on HD data, first, HFB responses were extracted for speech and non-speech fragments. The data were averaged per fragment and compared across speech and non-speech fragments using t-tests. Next, we calculated Spearman cross-correlations with sound envelope and cross-correlations with the STG electrodes. The procedures were identical to the ones carried out on HD data.

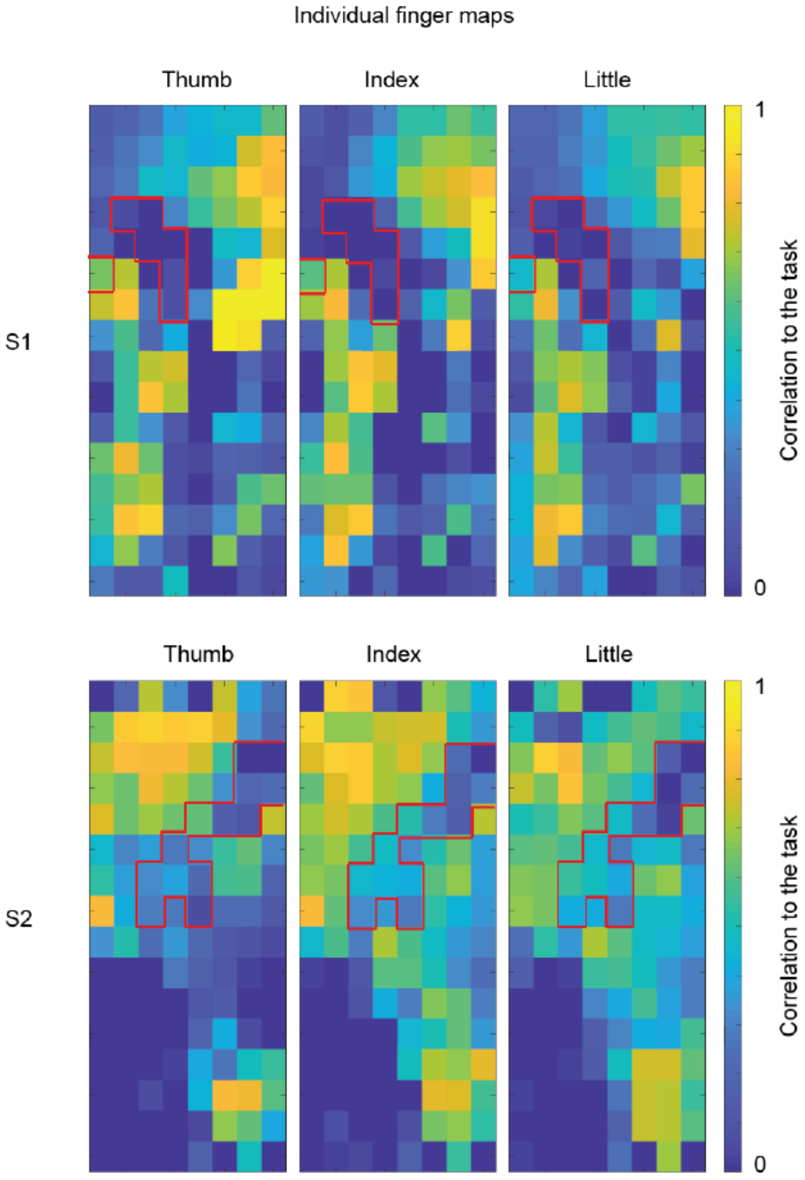
### Data and code availability

The data and custom code supporting the current study have not been deposited in a public repository due to the restrictions on public sharing of the patients' data but are available from the corresponding author on request.

# 4.5 SUPPLEMENTARY MATERIAL



**Supplementary Figure 4.1S.** Cortical maps (electrode grid plots) for individual speech articulators: lips, tongue, jaw and larynx. The red contour outlines electrodes involved in speech tracking (Figures 2a and 6a).



**Supplementary Figure 4.2S.** Cortical maps (electrode grid plots) for individual finger movements: thumb, index and little finger. The red contour outlines electrodes involved in speech tracking (Figures 4.2a and 4.9a).

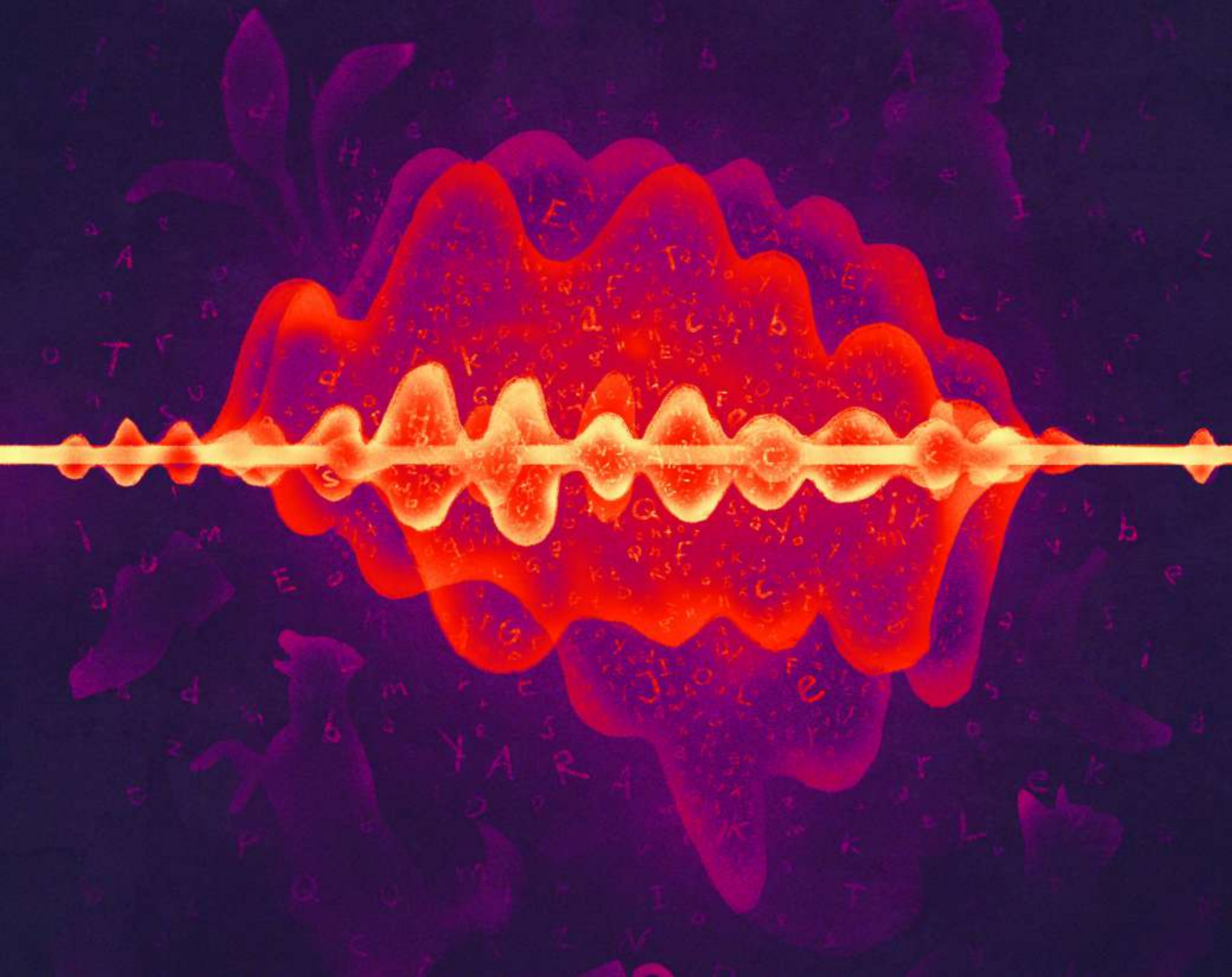
**Supplementary Table 4.1S. MNI-coordinates of the HD electrodes** with significant tacking of speech spectral envelope against the non-speech baseline (Figures 2a and 6a).

| Area | MNI coordinate | | | Patient |
|---|---|---|---|---|
| | x | y | z | |
| Left precentral gyrus | -46 | -10 | 61 | S1 |
| Left precentral gyrus | -48 | -9 | 58 | S1 |
| Left precentral gyrus | -51 | -9 | 55 | S1 |
| Left precentral gyrus | -41 | -6 | 63 | S1 |
| Left precentral gyrus | -44 | -5 | 61 | S1 |
| Left precentral gyrus | -39 | -2 | 63 | S1 |
| Left precentral gyrus | -45 | 4 | 57 | S1 |
| Right precentral gyrus | 57 | -7 | 51 | S2 |
| Right precentral gyrus | 59 | -6 | 49 | S2 |
| Right precentral gyrus | 54 | -7 | 54 | S2 |
| Right precentral gyrus | 56 | -5 | 52 | S2 |
| Right precentral gyrus | 51 | -6 | 57 | S2 |
| Right precentral gyrus | 54 | -3 | 53 | S2 |
| Right precentral gyrus | 56 | -1 | 50 | S2 |
| Right precentral gyrus | 50 | -4 | 58 | S2 |
| Right precentral gyrus | 44 | -5 | 62 | S2 |
| Right precentral gyrus | 46 | -4 | 60 | S2 |
| Right precentral gyrus | 48 | -2 | 59 | S2 |
| Right precentral gyrus | 42 | -3 | 63 | S2 |
| Right precentral gyrus | 44 | -1 | 61 | S2 |

chapter

5

# Cortical network responses map onto data-driven features that capture visual semantics of movie fragments



# CHAPTER 5. CORTICAL NETWORK RESPONSES MAP ONTO DATA-DRIVEN FEATURES THAT CAPTURE VISUAL SEMANTICS OF MOVIE FRAGMENTS[4]

Research on how the human brain extracts meaning from sensory input relies in principle on methodological reductionism. In the present study, we adopt a more holistic approach by modeling the cortical responses to semantic information that was extracted from the visual stream of a feature film, employing artificial neural network models. Advances in both computer vision and natural language processing were utilized to extract semantic representations from the film by combining perceptual and linguistic information. We tested whether these representations were useful in studying the human brain data. To this end, we collected electrocorticography responses to a short feature film from 37 subjects and fitted their whole-brain patterns using the semantic components extracted from film frames. We found that individual semantic components reflected fundamental semantic distinctions in the visual input, such as presence or absence of people, human movement, landscape scenes, human faces, etc. Moreover, each semantic component mapped onto a distinct functional cortical network involving high-level cognitive regions in occipitotemporal, frontal and parietal cortices. The present work demonstrates the potential of the data-driven methods from information processing fields to explain patterns of cortical responses, and contributes to the overall discussion about the encoding of the high-level perceptual information in the human brain.

[4] This chapter is based on Berezutskaya, J., Freudenburg, Z. V., Ambrogioni L., van Gerven, M. A., & Ramsey, N. F. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. *In review*.

## 5.1 INTRODUCTION

Semantic processing of audiovisual material is a topic of great interest in cognitive neuroscience. A lot of knowledge has been accrued with studies focusing on visual object processing, object categorization and processing of various semantic attributes of the visually perceived objects (Grill-Spector and Weiner, 2014; Kravitz et al., 2011; Martin, 2007). We have come to understand a lot about the ventral stream of visual processing in the brain (Grill-Spector and Malach, 2004; Mishkin et al., 1983), the object categorization ability of inferior temporal cortex (Haxby et al., 2001; Kiani et al., 2007; Kriegeskorte et al., 2008b) as well as the specific roles of the fusiform face area (Kanwisher et al., 1997), the parahippocampal place area (Epstein and Kanwisher, 1998), the motion processing temporal region (Tootell et al., 1995) and other areas involved in high-level visual processing (Grill-Spector and Malach, 2004; Hasson et al., 2003; Zeki et al., 1991).

Most studies address the topic with a reductionist approach, with carefully constructed tasks and stimuli to investigate a particular aspect of brain function. With new tools available to investigate high-dimensional phenomena, more holistic approaches become feasible. Complex brain mechanisms may be identified or characterized by mapping brain responses onto informational structures that are extracted from non-constrained sensory material. More concretely, one can utilize recent advances in computational modeling in different domains such as computer vision and natural language processing that are driven by extraction of high-level semantic relations directly from the unprocessed input, such as images (Krizhevsky et al., 2012; Szegedy et al., 2013) and texts (Bian et al., 2014; Devlin et al., 2018; Mikolov et al., 2013). Among the most recent and most powerful such advances are deep artificial neural network models. Not only do these models achieve unprecedented performance on solving complex tasks (for example, visual object identification or text classification), but they are also often claimed to make errors in semantic judgements that are similar to the errors humans make, thereby mimicking aspects of human cognition (Stallkamp et al., 2012; Taigman et al., 2014; Zhang et al., 2018).

Both of these factors contributed to an increasing interest towards these models in the cognitive neuroscience community. Many have investigated the internal representations extracted by artificial neural networks and used them to model, or ‘fit’, the brain activity associated with perception of images (Cichy et al., 2016; Güçlü and van Gerven, 2015a; Kietzmann et al., 2019; Yamins and DiCarlo, 2016), music (Güçlü et al., 2016) and speech sounds (Berezutskaya et al., 2017a; Kell et al., 2018) . Multiple studies showed that these models can fit the evoked cortical responses better than alternative representations of stimuli typically based on hand-engineered filters for low-level perceptual input, for example, Gabor filters for pixel or sound spectrogram input (Cadieu et al., 2014; Kell et al., 2018; Khaligh-Razavi and Kriegeskorte, 2014). Artificial neural network approaches to modeling language have also shown to correlate well with the cortical responses to language related stimuli (Carota et al., 2017; Güçlü and van Gerven, 2017; Zhang et al., 2019).

Yet many of these studies used stimuli that were constrained in one way or another to satisfy the specific question at hand. We think we can learn more about how the brain makes sense of large amounts of continuously perceived visual information if we examine brain signals in a natural context. In essence, here we present a case for feasibility, where we reduce the complexity of natural high-level visual material to key components that represent sematic information. This reduction is accomplished by mining visual information from the perceived input followed by modelling semantic and language information content. The extracted semantic components are used to model the cortical responses driven by high-level processing of perceptual information.

In the present study, we collected a large dataset, in which participants watched a short feature film while their brain responses were recorded with electrocorticography. In this study we focus on the visual aspect of the film. We passed the film frames first through a set of algorithms including an automatic visual object recognition system, and then through a language model to extract high-level semantic components. The resulting set of key components were then imposed on the ECoG data to assess cortical representation thereof. We found that 1) the reduction of the visual data complexity yielded principal components that are readily interpretable in terms of semantic concepts, and 2) these components were represented in brain signals and distributed functional cortical networks. Our approach shows that processing of meaning in the visual stream of the film maps onto brain regions associated with corresponding functionality: lateral fusiform gyrus for processing faces, parietotemporal network for processing movement, lateral occipital complex for processing objects, scenes and landscape shots. We believe this approach has potential for deepening our understanding about how the human brain extracts meaning from visual information presented in an unconstrained natural context.

## 5.2 RESULTS

We investigated whether the visual semantic information obtained from a feature film through a bottom-up computational approach could be informative in explaining the associated neural responses. First, we developed a semi-automatic bottom-up approach to extract principal *semantic components* that captured high-level conceptual information in the film's visual stream. The extracted components proved to reflect fundamental semantic differences between film frames, such as presence or absence of people, motion versus static landscape frames, human faces versus human bodies. Next, the semantic components were used to model the neural data collected from 37 subjects during a film-watching electrocorticography (ECoG) experiment. The model fit showed significant accuracy for predicting the brain activity, with the prediction accuracy peaking at 320 ms after frame onset primarily in occipitotemporal, parietal and inferior frontal cortices. Further analyses showed that the brain areas with significant accuracy could be subdivided into distinct cortical networks, each engaged in processing of specific semantic components in the film's visual stream of information.

### A semi-automatic bottom-up approach to obtain vectors of visual semantics from film frames

In order to extract semantic meaning, we combined the advanced technologies in various fields (deep learning, computer vision, natural language processing) to obtain visually driven semantic representations that can be used to study the neural responses. For this, we devised a pipeline that allowed us to combine recently published algorithms and obtain semantic representations.

Our semi-automatic pipeline contained three stages: the visual object recognition stage (I), the language model stage (II) and the dimensionality reduction stage (III, **Figure 5.1a**). In stage I we employed an artificial neural network model called *Clarifai* (www.clarifai.com), a state-of-the-art commercial computer vision model, to obtain labels of the objects and concepts present in each frame. *Clarifai* processes raw pixel information and generates the most likely concept labels together with their probabilities. The outputs of this model are referred to as *concepts* rather than *objects* because the system is capable of recognizing not only physical items in an image but also emotions, qualities, actions and some abstract concepts. The network used a preset dictionary of 5,000 concept labels and generated 20 concept labels per image frame. Despite the overall remarkable quality of the concept recognition system, we performed a manual check on the extracted labels and adjusted all incorrect assignments. Hence, we call our pipeline semi-automatic.

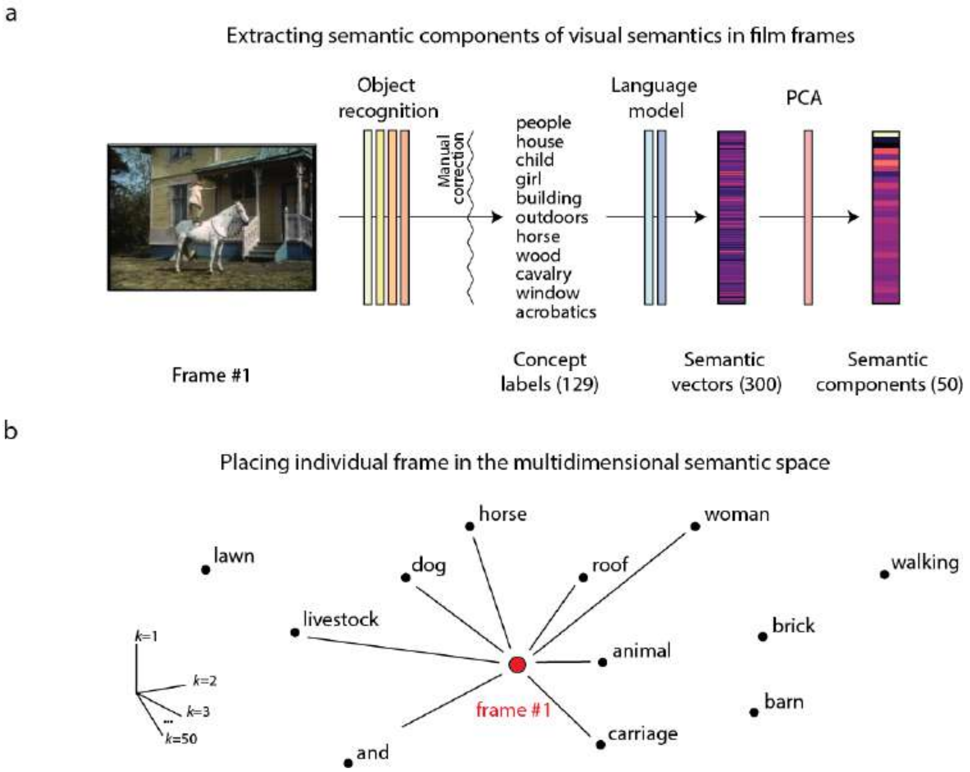


**Figure 5.1. Extraction of semantic components from the visual stream of the film. (a)** Semi-automatic pipeline for extraction of the semantic components. First, a film frame is passed through an automatic visual concept recognition system (*Clarifai*) to extract concept labels. Then, the extracted labels are passed though a language model (*fastText*) to obtain 300-dimensional semantic vectors, or word embeddings. The semantic vectors are averaged over all labels assigned to one frame, resulting in one averaged semantic vector per frame. The dimensionality of the vectors is further reduced by applying a principal component analysis to the averaged semantic vectors. The final result is a set of 50-dimensional semantic components that are used further to model the neural responses. **(b)** Example of how averaging of all concept labels per frame affects the semantic representation. Each word in the language model (*fastText*) can be seen as a point in a 300-dimensional semantic space. Neighboring words are assumed to capture similar semantics. Averaging in this space results in a new point that is placed in a neighborhood of all the words that are being averaged. Averaging has a capacity to represent combined complex meaning. In this example, the new point is in between the individual words 'horse', 'carriage' and 'roof', thus combining the meanings of these words together.

In principle, we could have used the extracted concept labels directly to model neural activity, as has been done before (Huth et al., 2012). However, our intention was not to limit the study to a preselected number of concept or object labels. Instead, we aimed to exploit the semantic relationships between different labels, including those absent in the present stimulus material, such that similar labels would get similar representations and dissimilar labels would get dissimilar representations. To achieve this, we subsequently applied a language model that could enrich our concept space with complex semantic and language relations between concept words.

To some extent, capturing of the semantics of the visually perceived concepts can be achieved through the co-occurrence of concept labels. However, the label co-occurrence could also lead to false associations. For example, the fact that the main character ('Pippi') and her horse co-occur in many frames does not necessarily mean than they should have similar semantic representations. Likewise, the fact that another character (Mrs. Settergren) only appears in the breakfast scene does not mean she should be associated specifically with the meal. Moreover, use of a universal space of the language model renders the model more generalizable in that if one were to study a different film with a different set of present visual objects, one would be able to project them onto the same semantic space as the present film and compare the results.

To extract semantic vectors of each concept label in the frame (stage II of the pipeline), an external language model was applied, called *fastText* (www.fasttext.cc, Bojanowski et al., 2017)*.* It is a shallow artificial neural network (*skip-gram* model) that has been trained on a large number of texts to extract word embeddings, or semantic vectors, for each word in the language vocabulary. The model has been shown to capture semantic relationships between words (Mikolov et al., 2013). The idea behind the use of semantic vectors is not only to simplify computations but to create a multidimensional representational space where mathematical operations such as addition and multiplication should apply. We obtained semantic vectors for each concept in the frame using the pretrained *fastText* model, and averaged the vectors over all concepts in the frame. This resulted in one semantic vector of the visual information per frame. We also verified that projecting the averaged semantic vector to the *fastText* space resulted in a point that was between, and closest to, the multiple individual concept labels of the frame (see example in **Figure 5.1b**).

The dimensionality of the data was further reduced to 1) represent the semantic space of our film more adequately (given that the concepts in the film likely covered only a small fraction of the entire semantic space of the *fastText* language model), and 2) determine the components of the most variance across the semantic vectors. To achieve this, we performed a principal component analysis on the previously obtained averaged semantic vectors (stage III). We found that projection to a space with just 50 principal semantic components accounted for 99.95% of all variance in the original averaged semantic vectors. Thus, from the averaged *fastText* semantic vectors per frame we obtained a smaller number of highly representable semantic components, ranked according to the amount of semantic variance they explained.

### Capturing of fundamental semantic distinctions in the extracted semantic components

Whether it would make sense to seek a relationship between the extracted semantic components and the brain responses depended of the capability of the adopted approach to capture meaningful distinctions and bits of semantics that drove the difference between scenes and separate events in the visual stream of the film. We were particularly interested in exploring the space of the extracted semantic components and determining whether they captured interpretable semantic information.

We found that the first five principal semantic components explained ~70% of all variance, suggesting that most of the semantic variability was captured by five fundamental semantic components. Each of the remaining components accounted for less than 5% of variance in the data and most of them (34 out of 45) for less than 1%.

To examine the top five extracted semantic components, we ranked the frames with respect to their values along each semantic component. For each component we selected frames corresponding to the bottom and top 10% of values and visualized them (**Figure 5.2**). In addition, we computed histograms of concept labels associated with the selected frames. Thus, the bottom 10% of frames together with their label histograms exhibited low values along a specific component, and the top 10% exhibited high values.

We observed that the first semantic component (34% explained variance) represented presence or absence of people in the frame. The second component (17% explained variance) differentiated between human movement and non-human movement or general nature scenes. The third component (11% explained variance) differentiated between scenes with movement (including travel and walking) and static scenes. The fourth component (9% explained variance) reflected differences between scenes with landscapes (including houses, rooms and other spaces) and portrait-like scenes (person or animal). Finally, the fifth semantic component (5% explained variance) captured human faces and differentiated them from scenes with human bodies and frames without people.

To assess the relationship between the semantic components and the labels we performed post-hoc statistical testing by fitting a linear regression to predict the values along the five semantic components using the concept labels per frame. The fit was significant for each of the top five semantic components ($R^2 > .99, \mathrm{F}(128;\ 9{,}675) = 14{,}490, p \ll .001$) and yielded regression weights for each of the components. The highest and lowest weights corresponded well to the histograms of concept labels for the top and bottom 10%, respectively, in that the highest weights for the first semantic component were assigned to labels ‘people’, ‘man’, ‘girl’, ‘adult’ and the lowest weights – to ‘park’, ‘tree’, ‘animal’, ‘outdoors’. The highest weights for the second component were assigned to labels 'climb, 'acrobatics', 'music', 'agility' and the lowest weights – to 'wildlife', 'outdoors', 'people', 'dark', etc. Comprehensive lists of highest and lowest weights as well as the 2D visualization of the overall structure of the semantic components (based on a t-SNE projection, see Methods for details) are shown in Supplementary Material (**Figures 5.1S and 5.2S**).

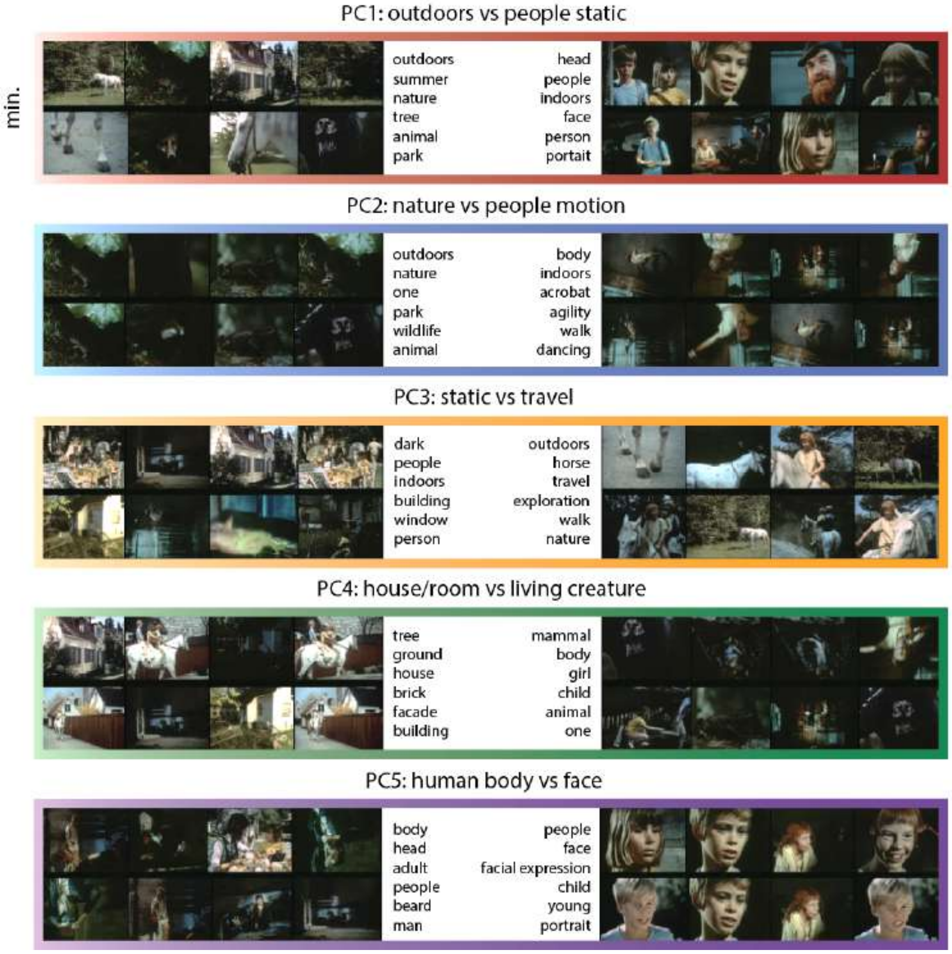


**Figure 5.2. Interpretation of the top five semantic components in the film's visual stream.** For each component we selected frames with bottom 10% (negative) and top 10% (positive) values. Thus, these selected frames express the range of semantic information captured in each component. The colors are unique identifiers of each components that are also used in later figures. The words associated with either bottom 10% or top 10% of frames per component are the most frequent concept labels present in these selected frames.

### High accuracy of predicting the neural responses based on the extracted semantic components

To assess whether the five semantic components mapped onto neural responses to the film (**Figure 5.3**), we fitted a neural encoding model (Holdgraf et al., 2017; Kay et al., 2008; Naselaris et al., 2012) where high frequency band (HFB) neural responses across all electrodes were predicted based on the extracted semantic components.

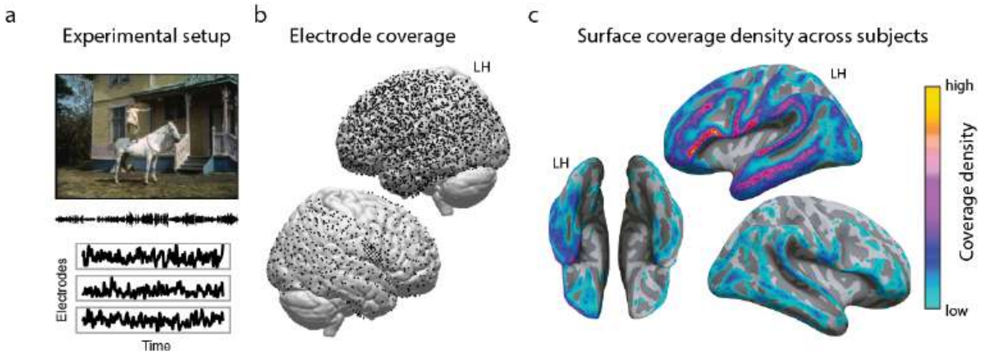


**Figure 5.3. Overview of the ECoG film-watching experiment. (a)** Experimental setup. The audiovisual film was presented to each participant while their neural responses were collected with ECoG electrodes. **(b)** Total electrode coverage over 37 participants. The panel shows a clear bias toward covering the left hemisphere. Each single subject electrode location is projected onto the MNI common brain. **(c)** The surface representation of the coverage density. Each single subject electrode location was assigned a value of one and projected to a regular grid on the average Freesurfer space. The resulting overlay shows the density of coverage over all individual participants. The panel shows the same bias for the left hemisphere as well as better overall coverage of inferior frontal and temporal cortices. Considerable coverage of inferior temporal and fusiform regions can be seen, whereas calcarine sulcus and lingual gyrus are covered quite scarcely. See Methods for more details about the projection to a uniform regular grid on the average surface.

To account for the complex multimodal nature of the film prior to fitting a model on the extracted semantic components, we first regressed out parts of the neural signal associated with the auditory stream of the film (**Figure 5.3S**). A ridge linear regression was then applied on the residual neural data to predict whole-brain HFB ECoG responses to the film using the semantic components, with Pearson correlation between predicted and observed HFB responses as performance metric. The reported performance was calculated in the held-out test set and was cross-validated (see Methods for details).

Given that semantic processing is a high-level cognitive process that has been shown to require a substantial delay relative to the stimulus onset, we also tested which time shift relative to the stimulus onset was best for predicting the neural responses (**Figure 5.4a**). A separate ridge linear regression model was fitted at every time shift within the range of -10 s to 10 s. The highest accuracy of prediction was found at a lag of 320 ms after the stimulus onset. Even though longer lags displayed progressively lower prediction accuracy, multiple lags around the stimulus onset lead to a high accuracy of the fit. This was most likely due to the high autocorrelation of the semantic data (**Figure 5.4S**).

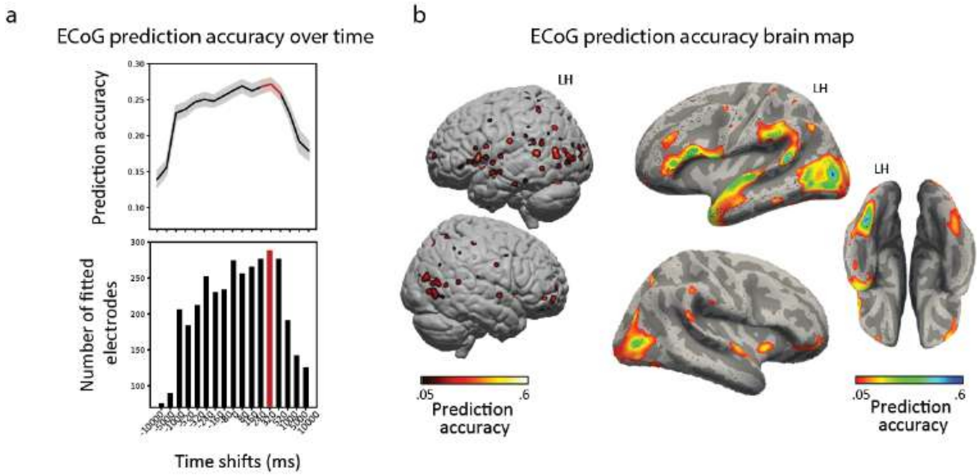


**Figure 5.4. Fitting the neural responses based on the extracted semantic components. (a)** Prediction accuracy (measured as Pearson correlation between predicted and observed HFB responses) for various time shifts of the semantic components. Prediction accuracy was cross-validated over five folds. Testing different time shifts was done to account for a delay in neural processing of the high-level visual information. The top panel shows the mean prediction accuracy for all significant electrodes ($p < .001$, Bonferroni corrected for the total number of electrodes) per time shift. Shading represents standard error of the mean. The bottom panel shows the number of electrodes with a significant fit per time shift. The red bar indicated the model with top average prediction accuracy as well as highest number of electrodes with a significant fit, which corresponds to a time shift of 320 ms. **(b)** Cortical map of the prediction accuracy for the time shift of 320 ms ($p < .001$, Bonferroni corrected for the total number of electrodes). Left panel shows a volume-based plot. Each cross-validated prediction accuracy value was assigned to the center coordinate of the corresponding electrode and projected to the MNI common space. Individual electrode locations were normalized to the MNI space using subject-specific affine transformation matrices obtained with SPM8. For the visualization purposes a 2D Gaussian kernel (FWHM = 8 mm) was applied to the coordinate on the MNI brain volume corresponding to the center of the electrode, so that the projected values (e.g. prediction accuracy) faded out from the center of the electrode toward its borders. Right panel shows a surface-based plot of the cross-validated prediction accuracy at time shift of 320 ms. It shows the same prediction accuracy values as left panel but projected on the surface for a better display of the fit in lateroocipital and fusiform cortices. See Methods for more details about the projection to a uniform regular grid on the average surface.

Considering the cortical map of the prediction accuracy, the best accuracy was achieved for occipitotemporal, parietal and inferior frontal cortices (**Figure 5.4b**). The prediction accuracy ranged from $r = .13$ to $r = .53$ (cross-validated Pearson correlation between predicted and observed HFB responses, 246 electrodes).

### Highly specialized cortical networks triggered by individual semantic components

Given that prediction accuracy was high in a number of brain regions bilaterally, we wondered whether all of the observed regions were involved in semantic visual processing to the same extent or whether there was a degree of specialization for some brain regions in responding to specific semantic components.

Investigation of the $\beta$-weights of the linear encoding model at the optimal time shift of 320 ms showed that distinct cortical networks were engaged in processing of the visual semantics of the film. Specifically, in order to identify groups of electrodes with similar $\beta$-weights across 50 components, and thus similar encoding of the semantic components, we applied a clustering approach to all $\beta$-weights. As a result, we identified a number of clusters, each characterized by a distinct cortical network as well as a ratio of contribution of the top five semantic components to the activation of that cluster (**Figure 5.5**). Thus, we observed for instance that cluster 1 was comprised of electrodes in lateral fusiform gyrus, and the semantic components encoding human presence (first semantic component) and human faces (fifth semantic component) contributed most to its activation time course.

Several cortical networks (clusters 2, 3 and 5) responded to the semantic components capturing motion. These networks comprised electrodes in supramarginal, precentral and posterior middle temporal gyri. The cortical network in cluster 2 seemed to be more specific to hand movement and object interaction (second semantic component), whereas the cortical network in cluster 3 seemed be more related to encoding of facial movement and facial expressions (combination of first, second and fifth semantic components). Cluster 5 appeared to be less sensitive to human presence in the frames but responded to biological motion and the semantic component of travel and transportation (third semantic component).

Cluster 4 was associated with presence of landscapes and static shots. Biological movement seemed to cause dips in its activation time course. Both clusters that included lateral occipital and dorsal parietal regions (clusters 4 and 5), favored frames without people presence and were deactivated by frames with human faces.

Not all of the cortical networks uncovered through clustering were as easily interpretable through the analysis of their distribution over the top five semantic components.

Remaining clusters with contribution from electrodes of many subjects still exhibited high anatomical consistency and included electrodes in inferior frontal gyrus (mostly pars opercularis) as well as posterior and anterior sites along the superior temporal gyrus (**Figure 5.5S**). It is unlikely that these cortical networks can be connected to tracking of the auditory information due to our extensive control for interaction between auditory and visual streams of the film. In addition, the activation time courses of these clusters do not seem to follow the block design, nor do they exhibit significantly larger correlation with the audio envelope compared to other clusters.

### Control for the low-level visual features

Because the semantic components we used were based on the output on the visual concept recognition model trained on raw image data, we aimed to implement some form of verification that the reported results were indeed due to the semantic processing rather than processing of lower-level visual features. First, we were able to confirm that pixel-level and Gabor-based visual features were dissociated from the extracted semantic components (two-sided Wilcoxon signed-rank tests: $\mathrm{Z} = -3{,}118, p \ll .001$ for pixel values and $\mathrm{Z} = -5{,}928, p \ll .001$ for Gabor features). Next, we used low-level visual features to predict the neural responses to the film with an intention to compare its prediction accuracy to the accuracy of the model that used the semantic components. Overall, we observed that the fit based on low-level features was considerably poorer ranging from $r = .13$ to $r = .26$ (cross-validated Pearson correlation between predicted and observed HFB responses, 17 electrodes). We think this was due to a number of factors. First, here we only had limited ECoG coverage in early visual cortex (**Figure 5.3c**). Second, the film-watching experiment did not have a fixation point in the center of the screen. Thus, unsurprisingly, using low-level visual features to predict the neural responses provided a significant fit in only a few electrodes outside of the early visual cortex (**Figure 5.6S**).

### Emergence of visual semantics from low-level visual features

Thus, we have shown that the observed prediction accuracy for the whole-brain responses was not due to low-level visual features. Since the visual concept recognition model was trained to extract high-level semantic information (concept labels) from the raw image data, somewhere along the layers of this deep artificial neural network, higher-level, semantic visual features were bound to emerge. We were curious whether we would be able to track the gradual buildup of these semantic representations throughout the visual concept recognition model and whether this buildup would be supported by the neural data. To this end, we investigated the relationship between the semantic components and each intermediate layer of a publicly available object recognition model called VGG16 (Simonyan and Zisserman, 2014), similarly trained to recognize objects in images by passing them through a set of convolutional layers. For simplicity, we only considered the *pooling* layers of the VGG16 model (see Methods for details). We found a

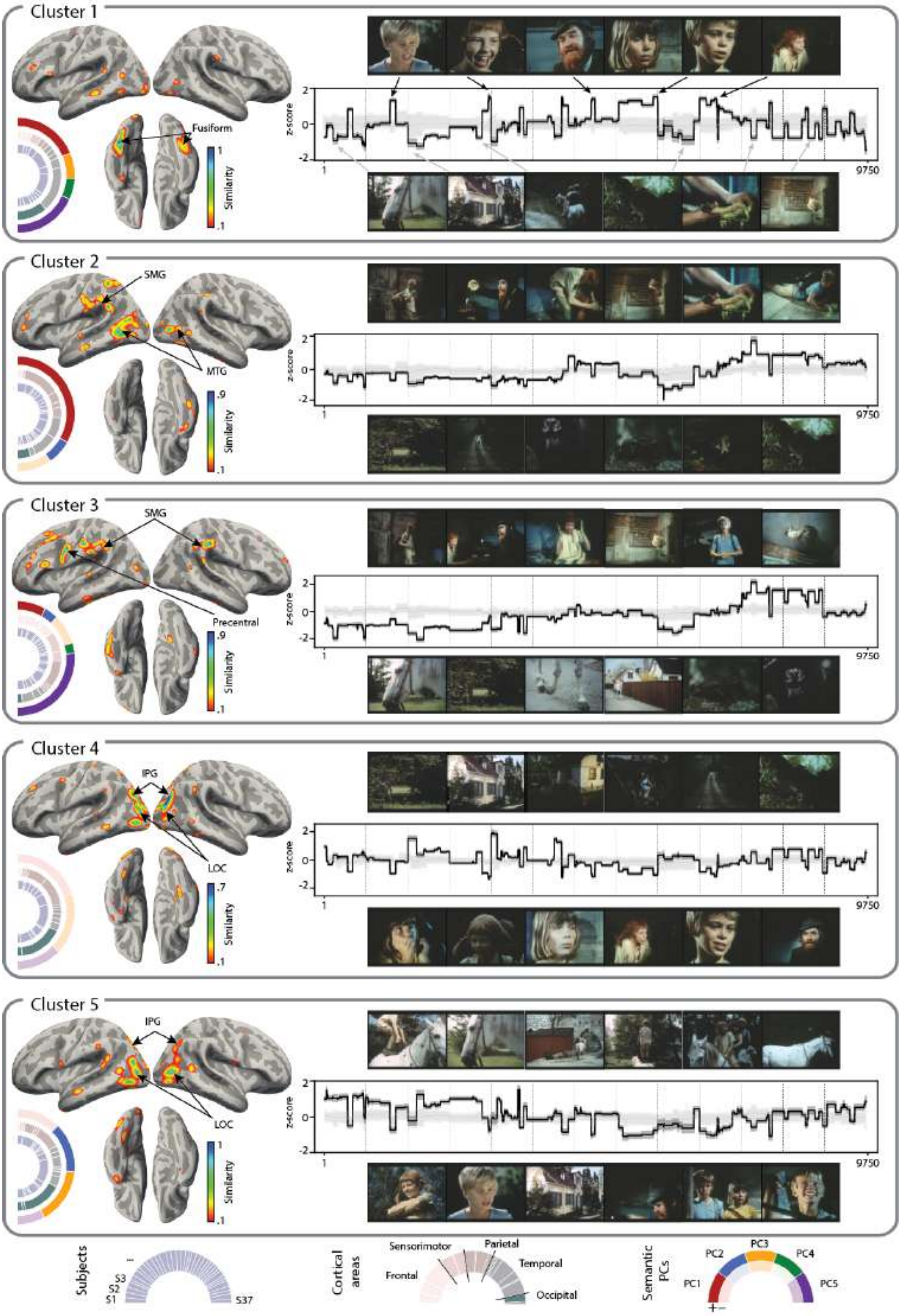

Cluster 1
Fusiform
Similarity
1
.1
z-score
2
0
-2
1
9750
Cluster 2
SMG
MTG
.9
.1
Cluster 3
SMG
Precentral
.9
.1
Cluster 4
IPG
LOC
.7
.1
Cluster 5
IPG
LOC
1
.1
Subjects
S1
S2
S3
...
S37
Cortical areas
Sensorimotor
Parietal
Frontal
Temporal
Occipital
Semantic PCs
PC1
PC2
PC3
PC4
PC5
+
−

**Figure 5.5. Distributed functional cortical networks associated with each individual semantic component.** Each cluster's profile contains diverse information including the cortical distribution of the electrodes contributing to that cluster (1), distribution over single subjects to show that all reported clusters included multiple subjects (2), distribution over cortical areas (3), distribution over the five semantic components (3), activation time course of the cluster (4) and frames that trigger and deactivate the cluster (5). The cortical maps are the surface-based projections of the similarity of each electrode in the cluster to the cluster exemplar. The similarity is measured as Pearson correlation of the $\beta$-weights across all electrodes. The $\beta$-weights are vectors of regression coefficients over 50 semantic components produced by the ridge linear regression  fit. Distributions over subjects (inner pie charts) show proportion of all electrodes per cluster that came from each single subject. None of the reported clusters was subject-specific, meaning that no more that 30% of electrodes per cluster came from a single subject. Distributions over cortical labels (middle pie charts) are color coded and only five groups of labels are highlighted: frontal, sensorimotor, parietal, temporal and occipital regions. The proportion of labels within each color coded region is still informative, such that in Cluster 1, for example, the largest contributing label is fusiform gyrus, which is a temporal region. It is not additionally color coded, but one can see that among the temporal regions there was only one with the largest contribution. Other times, contribution can be more equally distributed over multiple regions with the same color coded labels. Distribution over semantic components (outer pie charts) was calculated by performing a signed difference test on the top 10% of frames associated with peaks in the cluster's activation time course and the bottom 10% of frames associated with dips in the cluster's activation time course. Per cluster, the signed difference test (two-sided Wilcoxon ranked test) was performed by comparing values along each semantic component between 'peak frames' and 'dip frames'. The pie charts represent the test statistic, significant at $p < .001$ (Bonferroni corrected for the number of cluster and semantic components), adjusted for the decreasing percentage of explained variance from the first to the fifth semantic component (see Methods for details). The cluster activation time courses show the dot product of the semantic components and the $\beta$-weights of the exemplar of each cluster. Shading represents the standard error of the mean calculated on the dot product using $\beta$-weights of all electrodes of the cluster. Examples of frames associated with peaks in the cluster's activation are displayed above the activation time course. Examples of frames associated with dips in the cluster's activation are shown below the activation time course.

gradual increase in similarity between frame representations from the first to the last intermediate layer of the object recognition model (**Figure 5.6a**). Interestingly, the fit to neural data based on each individual intermediate layer of the object recognition model also showed a gradual increase in prediction accuracy (from the first to the last layer) together with the spread of the location of the best fitted electrodes from occipital toward temporal and parietal cortices (**Figure 5.7S**).

We then used the last intermediate pooling layer of the object recognition network (layer *pool5*, which we would expect is sensitive to complex patterns of object parts and general object shapes based on Güçlü and van Gerven, 2015a and Zeiler and Fergus, 2013) and compared the neural fit on *pool5* features to the fit on the semantic components. The difference in prediction accuracy between the fit on the semantic components and the fit on *pool5* features was observed for 30% of all electrodes favoring the fit on the semantic components ($r_{sem} - r_{pool5} > .1$). The overall difference in the whole-brain prediction accuracy was statistically significant: $\mathrm{Z} = 17.86, p \ll .001$, Bonferroni corrected for the number of electrodes, as assessed with a one-sided Wilcoxon signed rank test (**Figure 5.6c**). Electrodes in occipitotemporal, parietal and frontal cortices showed a better fit using the semantic components compared to the fit on *pool5* features (**Figure 5.6d**).

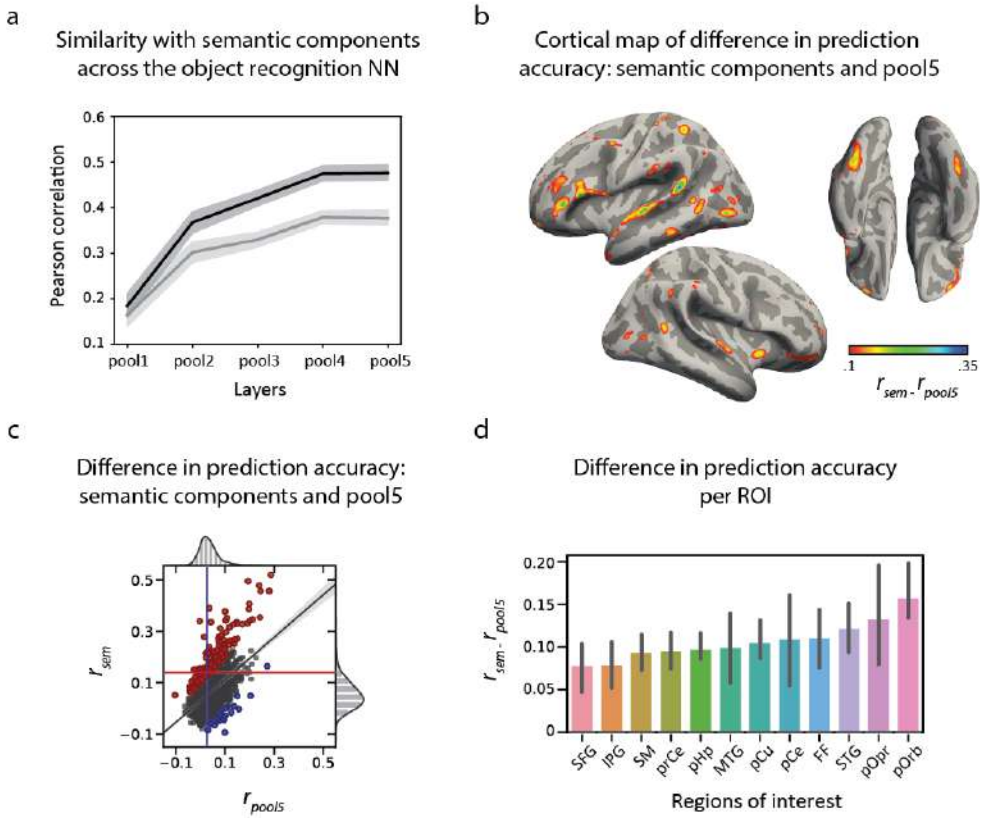


**Figure 5.6. Gradual emergence of the semantic information from low-level visual features. (a)** Similarity of frame representations between each intermediate layer of the automatic visual object recognition model and the semantic components used to fit the brain data. We used all so-called *pooling* layers of the object recognition model as intermediate layers. First layer (*pool1*) is placed quite early in the model (it is preceded by only two convolutional layers), whereas last layer (*pool5*) is located very deep in the model (followed by two dense layers and a probability output layer). The similarity with the semantic components is measured as Pearson correlation of all pairwise frame comparisons between each pooling layer and semantic components (see Methods for details). The shading represents 95th confidence interval based on the bootstrapping procedure (sampling of 1,000 frames 10,000 times). **(b)** Cortical map of the difference in prediction accuracy between the fit using features of the last pooling layer (*pool5*) and

the semantic components. **(c)** Scatter plot showing the difference in prediction accuracy between the brain fit using *pool5* ($r_{pool5}$) and the semantic components ($r_{sem}$). Each point represents a cross-validated accuracy per individual electrode. Red colored points denote electrodes with $r_{sem} - r_{pool5} > .1$, blue colored points denote electrodes with $r_{pool5} - r_{sem} > .1$. Red colored line shows median prediction accuracy over all electrodes with significant fit for the semantic components, blue colored line shows the same value for *pool5*. **(d)** Significant difference in prediction accuracy for individual regions of interest (ROI). Each bar shows a mean difference $r_{sem} - r_{pool}$ per ROI. Error bars represent 95th confidence interval based on the bootstrapping procedure (sampling of 50% electrodes per ROI 10,000 times).

Overall, these results indicated a gradual emergence of the semantic features from the low-level visual information in the object recognition model. At the same time, even the top layer of the object recognition model (*pool5*) provided an inferior fit of the neural data compared to the semantic components. The cortical areas with the overall significant difference in prediction accuracy were high-level cognitive processing regions including superior and inferior frontal cortices, temporal, parietal and motor regions. Only the occipital regions, which are specifically involved in visual processing, did not show a significant difference.

### Contribution of the language model to the extracted semantic components

Finally, having seen that the semantic components provided a better fit for the brain responses throughout high-level cognition areas compared to the top pooling layer of the visual neural network, we wondered whether the language model contributed to model accuracy.

To address this point, we fitted another ridge linear regression model using binary vectors of concept labels to predict HFB neural responses. We then compared the prediction accuracy with the model that used semantic components. We found that even though there did not seem to be a specific brain region where the accuracy was significantly better for the semantic component model, the model that used semantic components on average provided an overall better prediction accuracy across the cortex (one-sided Wilcoxon signed rank test: $\mathrm{Z} = 2.97, p < .05$, Bonferroni corrected for the number of electrodes with a significant model fit).

# 5.3 DISCUSSION

In the present study we show that the automatically derived semantic properties of the visual narrative in a feature film are captured in distributed cortical networks, each associated with distinct semantic components. In particular, we were able to combine recent advances in both visual object recognition and natural language processing to develop a semi-automatic approach to extract semantic components from the visual stream of the film. Modelling the associated neural responses on the basis of the semantic components resulted in significant prediction accuracy peaking at 320 ms after the frame onset, primarily in occipitotemporal, parietal and inferior frontal cortices. Investigation of the model weights showed that distinct cortical networks were engaged in processing of the visual semantics of the film with lateral fusiform gyrus processing faces, supramarginal, motor and posterior middle temporal regions processing movement, and lateral occipital regions processing complex static scene information.

## Fitting neural responses with semantic information derived through automatic processing of stimuli

Previous research on neural encoding models using deep neural network representations shows their potential in explaining brain activity. Multiple studies have reported that activation in both early and secondary visual cortex reflects similar representations of visual stimuli as the trained artificial neural networks (Cichy et al., 2016; Güçlü and van Gerven, 2015a; Yamins and DiCarlo, 2016). High-level semantic distinction and object representation have been traditionally associated with inferior temporal and fusiform cortices (Chao et al., 1999; Haxby et al., 2001; Kanwisher et al., 1997; Kiani et al., 2007), and it has been shown that activation patterns in these regions exhibit similarity with representations learned by top layers of object recognition neural networks (Güçlü and van Gerven, 2015b; Khaligh-Razavi and Kriegeskorte, 2014).

Language models that extract semantic properties of words by learning their co-occurrence patterns have also been successfully related to neural data through either encoding (Güçlü and van Gerven, 2017) or decoding (Abnar et al., 2017; Pereira et al., 2018) approaches. Exploiting the co-occurrence patterns to study encoding of meaning in the brain has proven fruitful regardless of whether the co-occurrence patterns were extracted through an artificial neural network (Güçlü and van Gerven, 2017; Pereira et al., 2018; Zhang et al., 2019) or simpler corpus-based approaches (Carota et al., 2017; Mitchell et al., 2008), underlining the importance of contextual information in semantic representation.

Here, an important step forward from this impressive previous work is the combination of the representations from different domains for predicting brain responses. We combine advances in both visual object processing and natural language processing to obtain rich semantic representations of the film's visual stream, that are informed by a language model. We find that this combined approach provides a better fit for whole-brain neural responses to a complex audio-visual narrative, than concept labels alone, even after having manually corrected their assignment.

### Gradual emergence of the semantic representations from low-level visual input

One of the main reasons why artificial neural network models are so interesting from the neuroscience point of view is their ability to extract hierarchical representations capturing transitions from low-level raw input data, such as images, to high-level semantic distinctions.

Previous research in visual neuroscience reported that increasingly complex representations learned by an image-based deep learning model, translated into a gradient along the cortical areas involved in visual object recognition. Along this gradient, responses in early visual cortex were fitted best by representations in early layers of the deep learning model, whereas responses in fusiform and inferior temporal cortex were fitted best by representations in later, more object- and category-specific layers of the deep learning model (Cichy et al., 2016; Güçlü and van Gerven, 2015b; Khaligh-Razavi and Kriegeskorte, 2014).

The present study corroborates the evidence that high-level semantic representations emerge gradually throughout the deep artificial neural network, and that this gradual shift maps onto neural data as well. That is, larger similarity of the later layers of the visual object recognition model with the semantic components was reflected in a better fit to the brain responses in high-level areas including occipitotemporal, parietal and some frontal regions. Our approach seems to provide a better fit compared to the top layer of the visual object recognition model in many high-level cognitive areas. This once again underlines the contribution of the use of manually corrected labels combined with a language model that captures semantic distinctions between the labels.

### Distributed functional cortical networks each encoding an individual semantic component

A previous attempt at combining the visual concept recognition model with a context-based language model has been made by Güçlü and van Gerven (2017). However, one of the strengths of the present study is the focus on exploring the semantic space of the extracted high-level representations, and encoding of individual semantic components in the neural responses. Thus, we show that the semantic distinctions observed along each of the top principal semantic components are associated with activation of specific functional cortical networks. The ability to uncover this mapping and interpret the cortical activity through processing of variance along distinct continuous semantic dimensions is the main contribution of this work.

The adopted approach separated the functional cortical networks associated with visual perception of human movement, human faces, general movement, landscapes and static scenes. The involvement of the lateral fusiform gyrus in processing of human faces has been reported in a large body of previous research (Haxby et al., 2001; Kanwisher et al., 1997), focusing on its role in identity perception (Haxby and Gobbini, 2011; Leopold and Rhodes, 2010). Interestingly, here the human faces seemed to be part of a semantic dimension, with nature, landscape and movement frames on one end and human faces on the other. This finding resonates with theories on existence of a so-called animacy continuum that drives difference in the neural responses to various visual stimuli (Connolly et al., 2012; Huth et al., 2012).

Processing of human movement in our study shows involvement of sensorimotor cortex, supramarginal gyrus and middle temporal region, or MT. The MT region is the classical area reported in perception of motion in general, whether it is movement of humans, objects or dot patterns (Gaglianese et al., 2017; Tootell et al., 1995; Wheaton et al., 2004). However, the other two regions are reportedly involved in a more abstract level of motion perception. Sensorimotor cortex (pre- and postcentral gyrus) has been implicated in visual perception of human action (Dushanova and Donoghue, 2010; Fadiga et al., 2005). Supramarginal gyrus with an extension to anterior intraparietal sulcus has been reported to be involved in perception of spatial relations as well as complex human and animal body movements (Böttger et al., 2010; de Gelder et al., 2015), and of hand movement and interaction (Grosbras et al., 2012; Wheaton et al., 2004).

Another functional cortical network that emerged from this analysis is a combination of the lateral occipital cortex with inferior and superior parietal gyri. It is associated with processing of scenes, places, salient regions and possibly of multiple objects in general (Dilks et al., 2013; Grill-Spector et al., 2001; Stanley and Rubin, 2003). Involvement of the parietal regions indicates simultaneous encoding of the spatial relations of the objects within a scene (Goodale and Milner, 1992; Mishkin et al., 1983). In the present work coordination of this network with the MT region occurs during perception of the movement in a scene or through a specific place, consistent with the view on integrative (form-motion) function of lateral occipital-temporal cortex (Beauchamp, 2005).

### Using continuous semantic components to map the patterns of the neural activity

The holistic approach in the current study complements what has been learned from research using traditional approaches to studying visual semantics and object categorization in controlled (Cichy et al., 2014; Haxby and Gobbini, 2011; O'toole et al.,

2005) and naturalistic (Bartels and Zeki, 2004; Hasson et al., 2004; Lahnakoski et al., 2012) experimental paradigms. But, rather than address hypotheses by formulating specific questions and constraining the experimental paradigm, the current bottom-up analysis of data obtained in a natural context maps deep structures of visual input onto neural activity. Instead of looking to decode pre-selected discrete semantic categories we focused on the variance along individual orthogonal semantic axes (Huth et al., 2016, 2012; Zhang et al., 2019) extracted from the raw input itself. The fact that the categories that have traditionally been investigated, such as faces, places, movement and body parts turned out to be the principal components in the visual domain of the feature film, underscores the usefulness of a more holistic approach for investigating neural substrates of attribution of semantic meaning to visual input. As such, the presented findings constitute an indication of the largest contrasts of semantic features that constitute a key dimension in the visual input that cortical networks respond to, which may shed some light on how the brain attributes semantic meaning in a natural situation.

### Limitations

The present study has a number of limitations. For instance, our interpretation of the cortical patterns supporting semantic processing is limited by the ECoG coverage. Not all cortical regions were sampled uniformly by ECoG electrodes and activity in deep and folded regions was not recorded. Yet, we do not claim to explain full-brain neural activity underlying semantic processing, but rather reveal that high-level properties of the visual stream of the film explain part of the variation of cerebral responses in distributed cortical networks. Despite the limitation in coverage, the number of ECoG participants (37 subjects) is relatively large for an ECoG study and the reported results are likely to generalize across individuals.

Another limitation lies in the experimental material. Being part of the battery of standard clinical tasks, the experiment did not use material of a full feature film, but comprised a reduced subset of 30 second excerpts of a feature film (in total, 6.5 minutes long), edited together for a coherent story. Such a stimulus accommodated the main purpose of the clinical task to compare the speech and music sound blocks of the film (30 s each). We minimized the effects of this stimulus manipulation by regressing out the block structure and the auditory envelope from the cortical responses. At the same, this manipulation might have still affected our data and overall, such task structure resulted in limited amount of data compared to using a full feature film material.

We also observed that the visual stream of the film was characterized by high consistency of semantic information in consecutive frames. Logically, this makes sense as in a real world the information we perceive shares a lot of high-level abstract properties over consecutive time points resulting in high autocorrelation of the semantic properties of perceived input. Even though it appears to be an inherent feature of naturalistic semantic processing, it made the investigation of the dynamics of semantic processing rather limited.

Finally, the contents of any feature film in general contain material that people like to observe, and logically are likely to include people, actions, movement, various locations and so on. This could lead to an inherent bias in the type of semantic information that can be extracted from such material. It could also explain the correspondence of the semantic components extracted here with features investigated in isolation with traditional approaches (faces, movement, places, etc). More research with larger stimulus material is needed to estimate the bias and its effect on our understanding about the way semantic information is represented in the human brain.

## Conclusions

In the present study, we combined advances in computer vision and natural language processing to automatically extract complex semantic information from the visual stream of a short feature film. When fitted to predict whole-brain neural responses, these continuous semantic features triggered activation of distinct functional cortical networks, each associated with an individual semantic component of the visual narrative. These results underscore the potential of computational models that extract high-level semantic information from input data to offer insight about how the human brain processes visual information and forms semantic representations of the perceived world.

## 5.4 MATERIALS AND METHODS

### Film stimulus

For the film-watching experiment we used a 6.5 minute short movie, made of fragments from "Pippi Langkous" (Pippi Långstrump, 1969) edited together to form a coherent plot. The task was part of the standard battery of clinical tasks performed with a purpose of presurgical functional language mapping. The film consisted of 13 interleaved blocks of speech and music, 30 seconds each (seven blocks of music, six blocks of speech). The movie was originally in Swedish but dubbed in Dutch.

### A semi-automatic bottom-up approach to obtain vectors of visual semantics from film frames

#### Automatic visual concept recognition model

In order to obtain concept labels per frame we first extracted all frames from the film's visual stream and converted them to image files. Then, a pretrained commercial deep artificial neural network *Clarifai* '*General'* (www.clarifai.com) was used to obtain the concept labels per frame image. Frame images with original RGB colors, 768 × 576 in size were input into the *Clarifai* concept recognition model. A preset dictionary of 5,000 unique concepts was used. The output of the *Clarifai* model contained 20 most likely concept labels per image (=frame) and a probability score per label. A total of 518 unique labels were assigned to frames in our image set.

The output of the visual concept recognition model was then manually corrected. First, we only considered labels with a probability more than 90%, and after making sure that this way we are not losing any unique relevant labels per frame, we discarded the labels with a lower probability. Then, we removed all the labels, which were incorrectly assigned to the images, for example, 'dog', 'piano', 'mirror', 'battlefield', 'zoo', etc. Then, we removed labels that we deemed irrelevant or difficult to interpret, for example, 'television', 'actor', 'abstract', 'surreal', 'insubstantial', 'illustration', etc. Finally, we restricted the list of all labels to nouns (such as 'people', 'nature', 'horse', 'food'), adjectives describing to a well-defined state or relation ( such as 'seated', 'equestrian', 'wooden') and some adverbs (such as 'together', 'indoors', 'outdoors'). We also kept various labels describing action (such as 'walk', 'travel', 'dance', 'climb', 'smile', etc.). However, most labels referred to objects present in the frame, for example, 'house', 'table', 'animal', 'rock' etc. The manual correction procedure resulted in a reduction of the list of unique labels to 129 labels. Finally, we manually checked that frames did not lack any relevant concept labels from the refined label list.

### Language model

We used a word embedding model to associate each film frame with a numerical representation that would capture the combined semantics of all the concept labels per frame. For this, we used a pretrained *fastText* model (www.fasttext.cc), which is the extended version of a skip-gram language model, trained by predicting the context of a target word.

We downloaded the semantic vectors that were learned by the *fastText* model when trained on English Wikipedia (Bojanowski et al., 2017). The downloaded material contained pairs of words and their corresponding numerical semantic vectors, or word embeddings. We looked up a corresponding vector for each label of each frame. The obtained vectors were of length 300, which means that the semantic space of the *fastText* language model was organized along 300 dimensions, each potentially capturing some relevant language information. Thus, words with similar meanings are represented by vectors with similar values along the 300 dimensions.

Then, per frame, we averaged all numerical semantic vectors (averaging over all labels per frame). This way we were able to obtain one semantic vector of length 300 per frame.

### Principal component analysis

The principal component analysis was performed on the averaged semantic vectors obtained in the previous step and pursued two goals. First, we aimed to reduce the dimensionality of the semantic vectors. Second, we were interested in a transformation of the semantic space that would uncover the dimensions of the most variance. We suspected that the semantic data used in the present study only captured a small set of the semantic distinctions between all words in the language model because of our focus only on the visually perceived concepts and due to the length and the narrative consistency of the film.

We set the number of principal components to 50 as this transformation provided a considerable reduction in dimensionality of the data while preserving over 99.9% of all variance.

## Interpretation of the extracted semantic vectors

To offer interpretation for the top five principal semantic components we focused on the examples at the minimum and maximum extreme ends per each component. We visualized the frames along with histograms of the concept labels per component.

Then, we performed a post-hoc statistical check by estimating an ordinary least squares fit for the semantic components based on the binary vectors of concept labels. Having observed that the fit was significant (under $p \ll .001$, Bonferroni corrected for the number of semantic components) for each of the top five semantic components, per semantic component we ranked the labels according to the regression weight values from most negative to most positive ones.

In addition, we visualized the multidimensional semantic space of 50 components using a 2D projection based on the *t-SNE* (Maaten and Hinton, 2008) algorithm (**Figure 5.2S**). T-SNE is a dimensionality reduction technique that projects high-dimensional data to a low-dimensional space suitable for visualization (for example, 2D). The algorithm first represents each data point in the high-dimensional space through a conditional probability distribution over its neighbors (*neighbor embeddings*). Then it computes the projection to the low-dimensional space that preserves these conditional probabilities (*neighbor embeddings*) by minimizing the Kullback-Leibler divergence between the high-dimensional and low-dimensional probability distributions. The *t-SNE* low-dimensional projection is considered one of the optimal techniques for high-dimensional data visualization that preserves relationships between data points at different scales (Maaten and Hinton, 2008). Here, we used the *scikit-learn* (Pedregosa et al., 2011) implementation of the t-SNE technique with default parameters (nearest neighbors = 30, learning rate = 200, metric = squared Euclidian distance, number of iterations = 1000).

## ECoG experiment

### Participants and procedures

All participants were admitted for diagnostic procedures with medication-resistant epilepsy. They underwent subdural electrode implantation to determine the source of seizures and test the possibility of surgical removal of the corresponding brain tissue. Research could be conducted between clinical procedures. All patients gave written informed consent to participate in accompanying electrocorticography (ECoG) recordings and gave permission to use their data for scientific research. The study was approved by the Medical Ethical Committee of the Utrecht University Medical Center in accordance with the Declaration of Helsinki (2013).

Thirty-seven patients (age 26±12, 24 females) participated in the film-watching experiment. Thirty patients were implanted with left hemispheric grids. Most patients had left hemisphere as language dominant, based on fMRI, Wada or functional transcranial Doppler sonography test (**Table 5.1**).

All patients were implanted with clinical electrode grids (2.3 mm exposed diameter, inter-electrode distance 10 mm, between 48 and 128 contact points); one patient had a high-density grid (1.3 mm exposed diameter, inter-electrode distance 3 mm). Almost all patients had temporal grid coverage and most had electrodes in frontal, parietal and motor cortices. The total brain coverage of the patients can be seen in **Figure 5.3**. Patient-specific information about the grid hemisphere, number of electrodes, and cortices covered is summarized in **Table 5.1**.

**Table 5.1. Electrode grid information for all participants.** Shown is information about the number of electrodes, grid hemisphere, covered cortices, handedness, and language-dominant hemisphere per patient. L, Left; R, right; F, frontal cortex; M, motor cortex; T, temporal cortex; P, parietal cortex; O, occipital cortex; fMRI, functional magnetic resonance imaging; fTCD, functional transcranial Doppler sonography.

| Patient | N of electrodes | Grid hemi-sphere | Cortices covered | Handed-ness | Language dominance |
|---|---|---|---|---|---|
| 1 | 96 | L | F, T, P | R | L (Wada) |
| 2 | 112 | L | F, M, T, P, O | R | L (fMRI) |
| 3 | 96 | L | F, M, T, P, O | R | L (Wada) |
| 4 | 104 | L | F, M, T, P | R | L (Wada) |
| 5 | 48 | L | F, M, T, P | L | L (fMRI) |
| 6 | 120 | L | F, M, T | R | L (Wada) |
| 7 | 112 | L | F, M, T, P | R | L (fMRI) |
| 8 | 64 | L | F, M, T | R | L (fMRI) |
| 9 | 112 | L | M, T, P, O | R | L (fMRI) |
| 10 | 64 | L | M, T, P | R | L (Wada) |
| 11 | 96 | R | M, T, P, O | L | R (Wada) |
| 12 | 88 | L | T, P, O | R | L (fMRI) |
| 13 | 112 | L | F, T, P | R | R (Wada) |
| 14 | 120 | R | F, M, T | R | L (Wada) |
| 15 | 96 | R | F, T, P, O | L | L (Wada) |
| 16 | 120 | L | F, T, P, O | not available | L (Wada) |
| 17 | 88 | L | F, M | R | L (fMRI) |
| 18 | 96 | L | F, M, T, P, O | L | L (Wada) |
| 19 | 64 | R | T, P, O | R | not assessed |
| 20 | 64 | L | F, M, T, P | R | L (fMRI) |
| 21 | 88 | L | F, M, P | R | L (fMRI) |
| 22 | 80 | L | F, M, T, P | R | L (Wada) |
| 23 | 128 | L | F, M, T, P, O | R | L (fMRI) |
| 24 | 80 | R | F, M, T, P, O | R | L (fMRI) |

| | | | | | |
|---|---|---|---|---|---|
| 25 | 64 | L | M, T, P, O | L | L (fMRI) |
| 26 | 64 | R | T, P, O | R | L (Wada) |
| 27 | 64 | L | F, M, T | R | L (fMRI) |
| 28 | 64 | L | F, M, T | R | L (fMRI) |
| 29 | 48 | R | F, M, T | L | R (fMRI) |
| 30 | 112 | L | F, M, T | R | R (fMRI) |
| 31 | 64 | L | F, M, T | R | L (fTCD) |
| 32 | 64 | L | F, M, T, P | R | L (fMRI) |
| 33 | 120 | L | T, P, O | R | L (fMRI) |
| 34 | 72 | L | F, M | L | R (Wada) |
| 35 | 72 | L | F, M | R | bilateral (fMRI) |
| 36 | 96 | L | F, M, T | R | L (Wada) |
| 37 | 80 | L | F, M, T | R | L (fTCD) |

In the film-watching experiment, each patient was asked to attend to the film displayed on a computer screen (21 inches in diagonal). The stereo sound was delivered through speakers with the volume level adjusted for each patient.

During the experiment ECoG data were acquired with a 128-channel recording system (Micromed, Treviso, Italy) at a sampling rate of 512 Hz filtered at 0.15–134.4 Hz. The film was presented using Presentation® (Version 18.0, Neurobehavioral Systems Inc) and sound was synchronized with the ECoG recordings.

## ECoG data processing

All electrodes with noisy or flat signal (visual inspection) were excluded from further analyses. After applying a notch filter for line noise (50 and 100 Hz), common average rereferencing was applied to all clinical grids per patient (and separately for the one high-density grid). Data were transformed to the frequency domain using Gabor wavelet decomposition at 1–120 Hz in 1 Hz bins with decreasing window length (four wavelength full-width at half maximum). Finally, high frequency band (HFB) amplitude was obtained by averaging amplitudes for the 60–120 Hz bins and the resulting time series per electrode were down-sampled to 25 Hz, which corresponded to the frame rate of the film. Electrode locations were coregistered to the anatomical MRI in native space using computer tomography scans (Branco et al., 2018; Hermes et al., 2010) and Freesurfer (http://surfer.nmr.mgh.harvard.edu/). The Desikan–Killiany atlas (Desikan et al., 2006) was used for anatomical labeling of electrodes (closest cortical structure in the radius of 5 mm). All electrode positions were projected to Montreal Neurological Institute space using SPM8 (Welcome Trust Centre for Neuroimaging, University College London).

## Brain visualization: volume and surface projections

For volume-based visualizations, we used electrode projections to the subject-specific anatomical volume obtained with the electrode localization tool (Branco et al., 2018). Then, individual electrode locations were normalized to the MNI space using patient-specific affine transformation matrices obtained with SPM8. For the visualization purposes a 2D Gaussian kernel (FWHM = 8 mm) was applied to the coordinate on the MNI brain volume corresponding to the center of the electrode, so that the projected values (e.g. prediction accuracy) faded out from the center of the electrode toward its borders.

For surface-based visualizations we used Freesurfer functions to project the volume-based electrode coordinates to the subject-specific anatomical surface. Then, these coordinates were projected to the subjects' common Freesurfer space for further visualization on the inflated surface.

However, we noticed that raw projections of the electrode center coordinates resulted in quite patchy visualizations, so we decided to apply some smoothing to all of the cortical overlays based on the electrode values. Because of the often occurring overlaps in electrode grids within and between subjects as well as irregularities of grid placement across subjects, in order to ensure a good result we imposed a regular uniform grid on the common Freesurfer surface and made projections of each individual electrode to the closest point of the regular grid. We used the uniform icosahedron grid of order 5 – *ico5*, distributed with the Freesurfer package. The smoothing was achieved by applying a Gaussian process algorithm (Rasmussen, 2003) to transform the values of each single subject electrode to the regular grid in the average subject space. Using this approach allowed for a projection of all values to a regular grid while taking into account the cortical distances between electrodes. Values coming from multiple electrodes to be projected onto the same regular grid coordinate were combined by considering their distances to that coordinate as well. We used the following Gaussian process implementation with an exponential kernel for the electrode distance matrices:

$$\mathrm{g} = \mathrm{K}_\mathrm{c}^\mathrm{T}(\mathrm{K}_\mathrm{s} + \eta \mathrm{I})^{-1}\mathrm{y} \tag{5.1}$$

$$K_c = \exp\left(-\frac{D_c^2}{2\sigma^2}\right) \tag{5.2}$$

$$K_s = \exp\left(-\frac{D_s^2}{2\sigma^2}\right) \tag{5.3}$$

where $D_s$ is a matrix of pairwise distances between electrodes of a single subject, $D_c$ is a matrix of pairwise distance between electrodes of a single subject and the points of the regular grid *ico5*, $\eta$ and $\sigma$ are the free parameters: $\eta$ is a noise parameter and $\sigma$ is the amount of smoothing in the exponential kernel. Vector y represents values to be projected from the single subject electrode space to the regular grid space. The distance matrices $D_s$ and $D_c$ are calculated as great circle distances on the average Freesurfer sphere surface:

$$d_{\mathrm{uv}} = r \ \cdot \mathrm{arcos}(\mathrm{u}^{\mathrm{T}}\mathrm{v}) \tag{5.4}$$

where $r$ is the sphere radius, and vectors u and v are the normalized coordinates of the points on the sphere. In case of subject's electrodes it's the vertex corresponding to the location of the center of the electrode and in case of the regular grid *ico5* it's the point on the grid.

Both matrices $D_s$ and $D_c$ are calculated using single subject electrode coordinates projected to the average Freesurfer sphere with Freesurfer resampling functions.

## High accuracy of predicting the neural responses based on the extracted semantic components

### Neural encoding model based on the semantic components

We used the previously extracted semantic components to predict the neural responses. A ridge linear regression model was employed. Pearson correlation between predicted and observed HFB responses in a held-out test set was used to evaluate model performance. The model performance was cross-validated using five-fold cross-validation. The values of the regularization parameter were determined using five-fold nested cross-validation. The reported correlations were significant at $p < .001$, Bonferroni corrected for the number of electrodes.

Because of the possible interaction between auditory and visual streams of the short film, we implemented a number of corrections in our model. First, because the auditory stream contained a block design with interleaved blocks of speech and music, we regressed out both the block design and the auditory envelope from both the semantic components and the HFB signal. In case of HFB signal, we first determined the optimal lag for regressing out block design and audio envelope through linear fitting at multiple lags (block design) or best lag of cross-correlation (audio envelope). Both were done independently and separately per each electrode. Cortical maps of regression to the block design and cross-correlation to audio envelope along with the histograms of lags and electrode labels are shown in Supplementary Material (**Figure 5.3S**).

Apart from fitting the regression on semantic components to the residuals of the regressions to block design and audio envelope, we also made sure that the test set per each cross-validation fold contained time points from multiple music and speech blocks. That is the test set of each cross-validation fold was constructed by concatenation of 6 second fragments across multiple music and speech blocks. This was done to avoid block specific effects in testing of each cross-validation fold.

### Testing various time shifts

Having finalized the details of the linear regression fit, we then performed the fit at multiple time shifts around the stimulus onset to determine the amount of delay in high-level semantic processing of the visual information. Per time shift, we analyzed the average model performance as well as the number of electrodes with a significant fit (at $p < .001$, Bonferroni corrected for the number of electrodes).

### Relation of the cortical networks to individual semantic components

Next, we aimed to investigate the relationship between brain responses and individual semantic components. In a linear regression model this relationship is reflected in the sign and magnitude of the regression coefficients, or $\beta$-weights. Thus, we focused on the $\beta$-weights of the neural encoding model fitted on the semantic components at the time shift that provided highest prediction accuracy (~320 ms). We selected the $\beta$-weights only of those electrodes, whose HFB responses were predicted significantly well by the model at $p < .001$, Bonferroni corrected for the total number of electrodes. The selected $\beta$-weights were averaged over five cross-validation folds and z-scored over electrodes per semantic component.

Then, an affinity propagation clustering (Frey and Dueck, 2007) approach was employed to find groups of electrodes with similar $\beta$-weight profiles across the semantic components. This clustering approach was used due to its non-parametric nature (no requirement to specify the number of clusters beforehand) as well as the ability to identify cluster *exemplars*, or data points representative of the entire cluster. We varied the value of the *preference* parameter but the main set of clusters (**Figure 5.5**) was found for all the clustering configurations. The present results are reported for the preference value equal to $\min(A) - 3$, where $A$ is the affinity matrix. Similarly, we used either Pearson or Spearman correlation coefficient for computing the affinity matrix, and found no significant difference. The reported results were obtained using Pearson correlations.

The described clustering configuration produced 13 clusters, however we only reported clusters that were non-subject-specific, i.e. they contained electrodes from at least one third of all subjects (12/37) and no more than a third of all electrodes in the cluster came from a single subject. Thus, we discarded six clusters: cluster 9 (7 different subjects; 65% electrodes came from a single subject), cluster 10 (7; 36%), cluster 11 (11; 37%), cluster 12 (4; 75%), cluster 13 (14; 42%) and cluster 14 (12; 70%).

For the remaining clusters we reported full cluster profiles containing information about the distribution over subjects, distribution over cortical regions, activation time course and the cortical projection map. The distribution over cortical regions was obtained by computing histograms over the cortical labels, associated with the location of the center of each electrode. The activation time courses were calculated as dot product between the semantic component values and the $\beta$-weights of the cluster exemplar, which is the electrode with the most representative cluster-specific $\beta$-weight profile.

To explore the relationship between each cluster's activation time course and individual semantic components we applied statistical testing to the frames that were associated with the peaks of the cluster's activation time course and the frames associated with its dips. In order to assess statistical significance of peaks and dips in each cluster's activation time course we shuffled cluster assignments (10,000 times) to obtain a baseline distribution of activation time courses per each cluster. Per cluster, we plotted its activation time course and highlighted the values corresponding to 2.5th and 97.5th percentiles of the baseline distribution (which corresponds to a two-tailed statistical threshold at p < .05). For subsequent analyses relating each cluster's time course to semantic components we only considered peaks and dips of the cluster's activation time course that were outside of the bulk of the baseline distribution. For peaks, we considered frames above the 97.5th percentile, for dips we considered frames below 2.5th percentile. Then, we selected the frames corresponding to top 10% of peaks and bottom 10% of dips in the cluster's activation time course. Per semantic component, we performed a two-sided Wilcoxon signed rank test to compare values along the semantic component in peaks and dips of the cluster's activation profile. The significance of the statistic was set at $p < .001$, Bonferroni corrected for the number of clusters and semantic components. The procedure was repeated for each cluster. The reported values of the statistic were corrected by the amount of variance each semantic component explained by dividing the statistic over the percentage of the explained variance. This was done to correct for the gradual decrease in magnitude of values along each semantic component due to its decreasing percentage of the explained variance.

### Control for the low-level visual features

As we sought to implement control for the confounding effect of the low-level visual features, we assessed the relationship between low-level visual features, the semantic components and the associated neural responses. First, we assessed the difference in the inner representation of the film frames in low-level features and the semantic components. As low-level features we used raw colored pixel values of the frame images (*pixel*) and Gabor features (*gabor*) configured to model retinotopic responses of complex cells in the early visual cortex of the human brain (Güçlü and van Gerven, 2014). Gabor features were extracted from the greyscale pixel values by passing them through a Gabor wavelet pyramid with predefined set of filter sizes and orientations. We followed the filter specifications used in Güçlü and van Gerven (2014).

Having extracted *pixel* and *gabor* features we computed the amount of similarity between each of them and the semantic components. For this, we computed pairwise correlations between all film frames using each type of representation: *pixel*, *gabor* and semantic components. Then, we assessed the difference between *pixel* and semantic components as well as *gabor* and semantic components using two-sided Wilcoxon signed rank tests.

Next, we assessed the difference in the prediction accuracy of the neural responses using *pixel*, *gabor* and semantic components. The ridge linear models using either *pixel* or *gabor* were fitted following the same procedures as previously described for the fit on the semantic components.

The difference in the prediction accuracy between the *pixel* model and the model using the semantic components was assessed using one-side Wilcoxon signed rank test on prediction values for all electrodes with significant performance in either model ($p < .001$, Bonferroni corrected for the number of electrodes). The difference between the *gabor* model and the model using the semantic components was assessed in the same way.

### Emergence of the visual semantics from the low-level visual features

Next, we assessed the relationship of the semantic components with the visual representations across the object recognition neural network. Because we previously used a commercial model to obtain the labels, we did not have access to its intermediate layer representations. Instead, here we used intermediate representations of another popular object recognition network, called VGG16 (Simonyan and Zisserman, 2014). VGG16 was pretrained to identify 1,000 object labels in input images. Due to a large network size, we limited our analyses to the representations of all the intermediate pooling layers ($n = 5$) of the VGG16 object recognition network. Thus, we passed every film frame through the pretrained VGG16 model and preserved only the representations of the five pooling layers. Due to a large number of frames, per filter of each pooling layer we applied 2D Gaussian smoothing to the frame representations and downsampled them along the image x and y dimensions.

Having obtained the frame representations per pooling layer, we calculated the amount of similarity between them and the frame representations based on the semantic components. This was done by first computing pairwise correlations between all frames per layer, which resulted in $l$ square matrices of size $f \times f$, where $l$ is the number of layers and $f$ is the number of frames. Same was done for the frame representations based on the semantic components. Then, per layer-specific matrix we took all its values in the upper triangle and correlated them with the upper triangle of the matrix based on semantic components. In essence, we performed a simplified version of the representational similarity analysis (Kriegeskorte et al., 2008a; Nili et al., 2014), where the semantic components were the target representation and the pooling VGG16 layers were the candidate representations and Pearson correlation was used as a similarity measure. For the statistical inference about the change in similarity with the semantic components across the pooling layers we performed bootstrapping. This was done by recalculating Pearson correlation between the frame similarity in semantic components and each pooling layer in random samples of 1,000 frames 10,000 times.

Then, we fit a ridge linear regression (following all the same procedures as previously described) on each of the intermediate VGG16 pooling layers to predict HFB ECoG responses. Similar to the fit using the semantic components, the prediction accuracy was cross-validated and projected on the brain volume. We also reported scatter plots showing the difference in prediction accuracy between the fit on pooling layers of VGG16 and the fit on the semantic components (**Figure 5.6c**, **Figure 5.7S**). The difference in the prediction accuracy was assessed using two-side Wilcoxon signed rank tests on prediction values for all electrodes with significant performance in either a layer-specific model or a model using the semantic components ($p < .001$, Bonferroni corrected for the number of electrodes).

### Contribution of the language model to the extracted semantic components

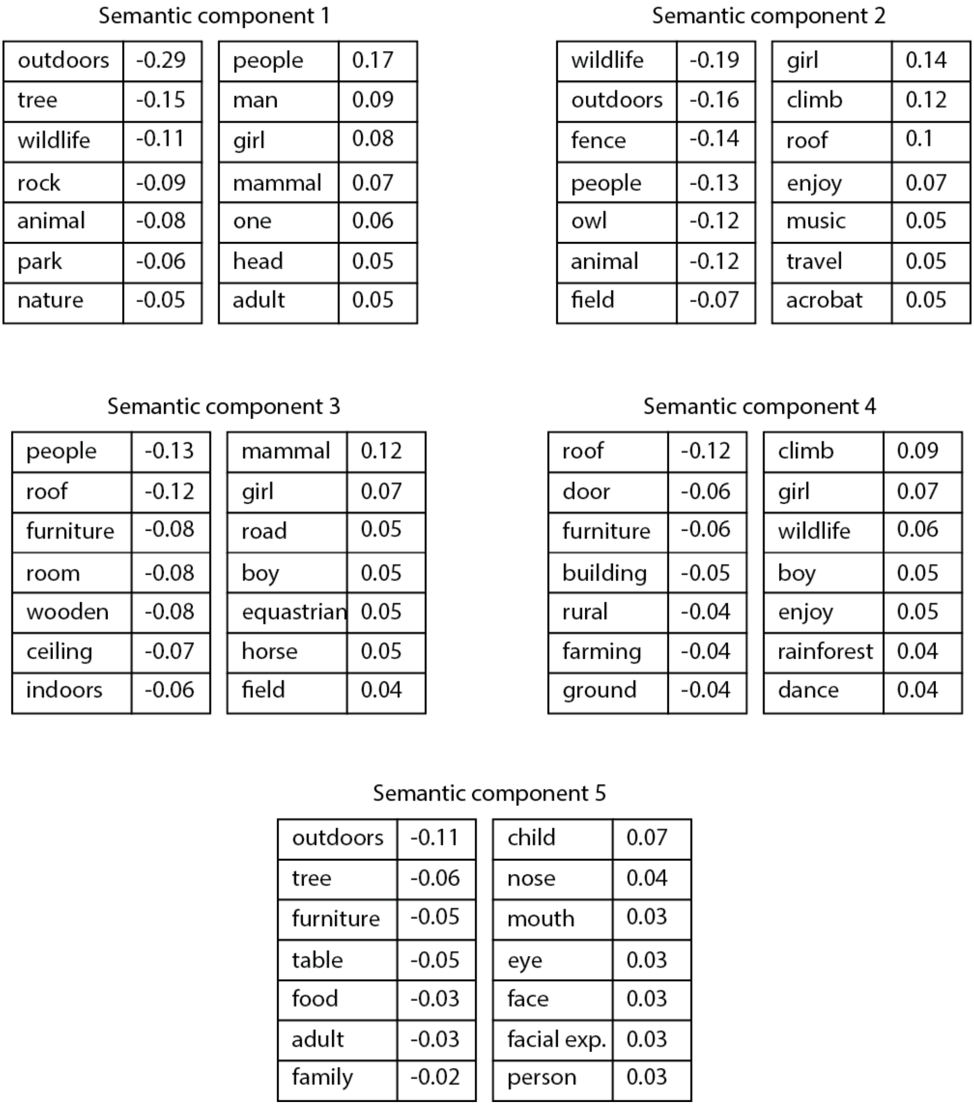
Finally, we analyzed the difference in prediction accuracy between the ridge linear model using the semantic components and the ridge linear model using the concept labels for prediction of the associated neural responses. The main difference between the two models was the usage of a language model in construction of the semantic components. In both models we used the manually corrected concept labels. Concept labels were represented as binary categorical vectors with each value in the vector corresponding to a specific label. Per frame, the value of zero corresponded to the absence of the corresponding concept in the frame image and the value of one corresponded to its presence. The ridge linear model was fitted following the same procedure as previously described for the fit on the semantic components except that the regression of the block design and the audio envelope was not applied to the categorical binary label vectors.

The difference in the prediction accuracy between the two models was assessed using one-side Wilcoxon signed rank test on prediction values for all electrodes with significant performance in either model ($p < .001$, Bonferroni corrected for the number of electrodes).

### Data and code availability

The data and custom code supporting the current study have not been deposited in a public repository due to the restrictions on public sharing of the patients' data but are available from the corresponding author on request.

## 5.5 SUPPLEMENTARY MATERIAL



Semantic component 1

| | | | |
|---|---|---|---|
| outdoors | -0.29 | people | 0.17 |
| tree | -0.15 | man | 0.09 |
| wildlife | -0.11 | girl | 0.08 |
| rock | -0.09 | mammal | 0.07 |
| animal | -0.08 | one | 0.06 |
| park | -0.06 | head | 0.05 |
| nature | -0.05 | adult | 0.05 |

Semantic component 2

| | | | |
|---|---|---|---|
| wildlife | -0.19 | girl | 0.14 |
| outdoors | -0.16 | climb | 0.12 |
| fence | -0.14 | roof | 0.1 |
| people | -0.13 | enjoy | 0.07 |
| owl | -0.12 | music | 0.05 |
| animal | -0.12 | travel | 0.05 |
| field | -0.07 | acrobat | 0.05 |

Semantic component 3

| | | | |
|---|---|---|---|
| people | -0.13 | mammal | 0.12 |
| roof | -0.12 | girl | 0.07 |
| furniture | -0.08 | road | 0.05 |
| room | -0.08 | boy | 0.05 |
| wooden | -0.08 | equastrian | 0.05 |
| ceiling | -0.07 | horse | 0.05 |
| indoors | -0.06 | field | 0.04 |

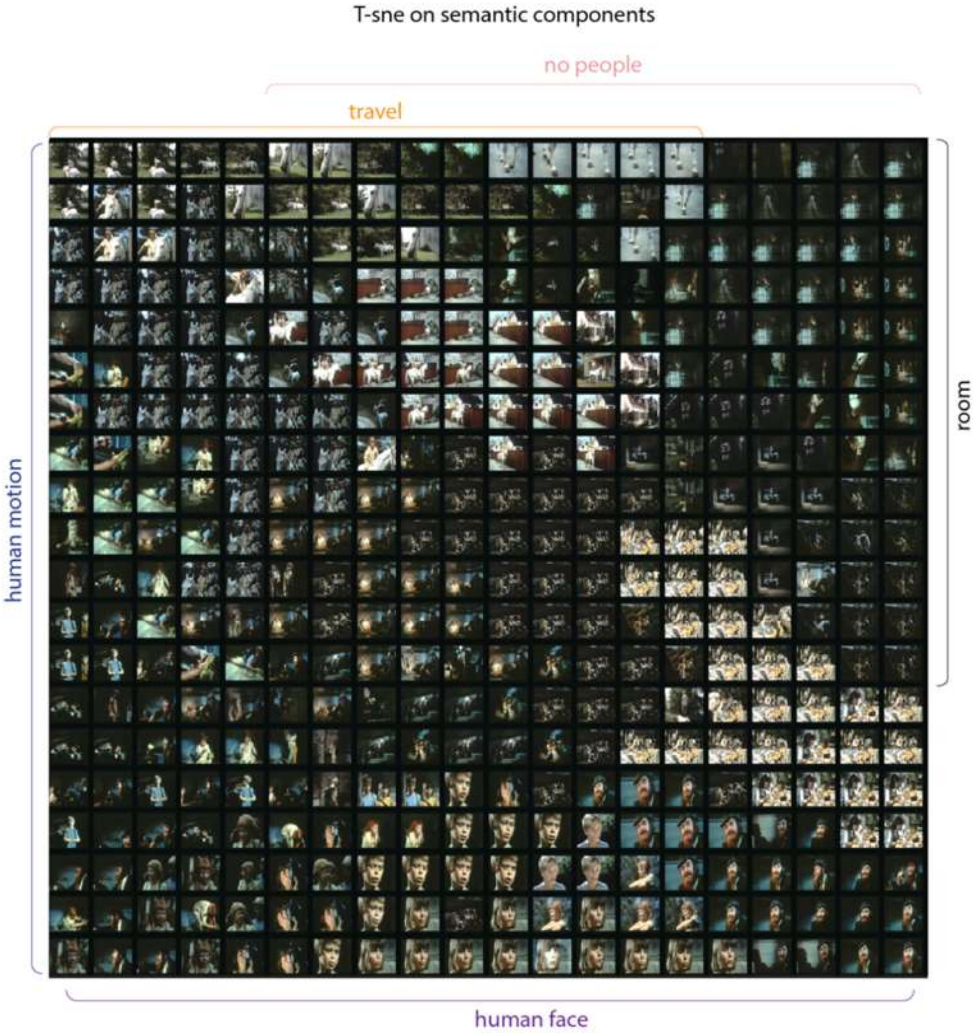
Semantic component 4

| | | | |
|---|---|---|---|
| roof | -0.12 | climb | 0.09 |
| door | -0.06 | girl | 0.07 |
| furniture | -0.06 | wildlife | 0.06 |
| building | -0.05 | boy | 0.05 |
| rural | -0.04 | enjoy | 0.05 |
| farming | -0.04 | rainforest | 0.04 |
| ground | -0.04 | dance | 0.04 |

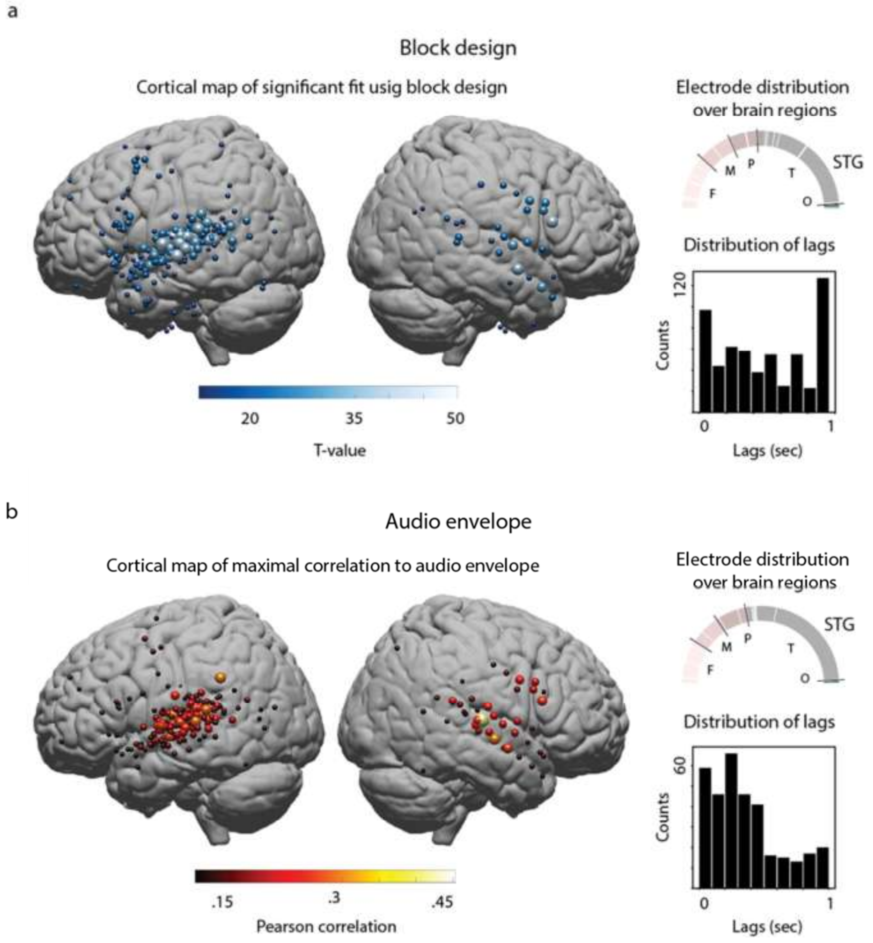
Semantic component 5

| | | | |
|---|---|---|---|
| outdoors | -0.11 | child | 0.07 |
| tree | -0.06 | nose | 0.04 |
| furniture | -0.05 | mouth | 0.03 |
| table | -0.05 | eye | 0.03 |
| food | -0.03 | face | 0.03 |
| adult | -0.03 | facial exp. | 0.03 |
| family | -0.02 | person | 0.03 |

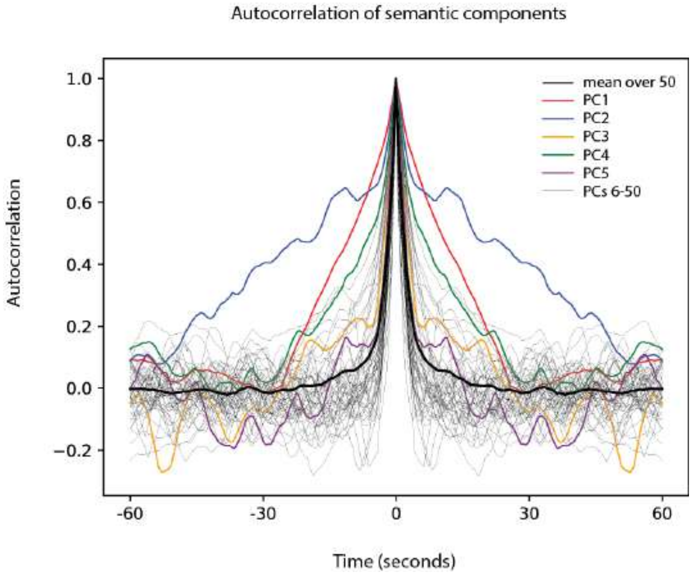
**Supplementary Figure 5.1S. Interpretation of semantic components.** Labels with largest absolute values of $\beta$-weights in the linear regression fitting concept labels to the semantic components.



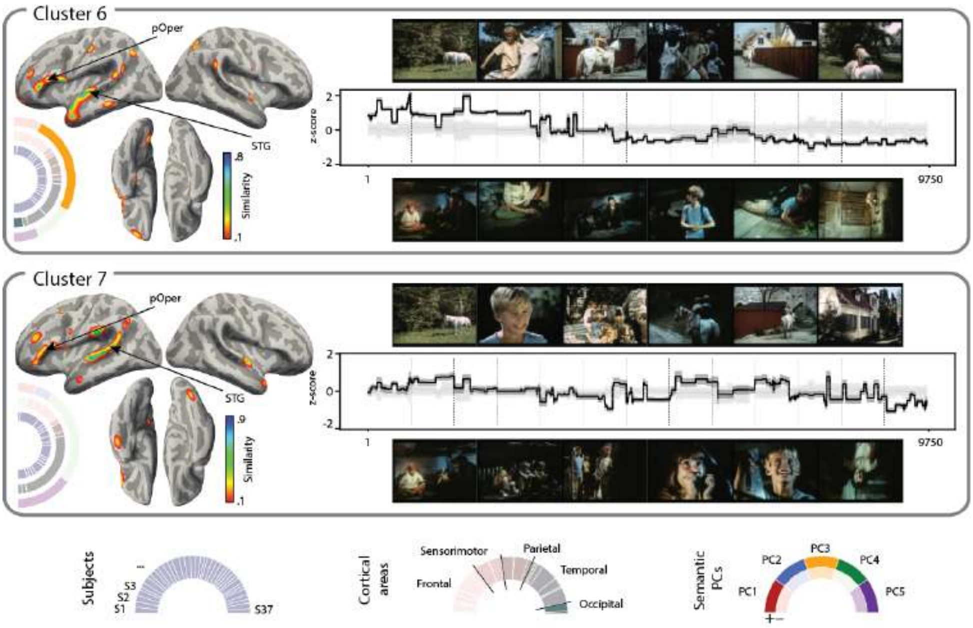
**Supplementary Figure 5.2S. T-SNE visualization of the frames based on their semantic component representation.** This visualization procedure projects a multidimensional space of semantic components (d=50) to a two-dimensional representation while preserving the distances between individual frames as much as possible. This projection offers a possibility to notice clusters of frames organized in distinct semantic groups: human faces at the bottom, human motion and object interaction on the left, closed spaces and no people present in top right corner, travel and human movement in top left corner.



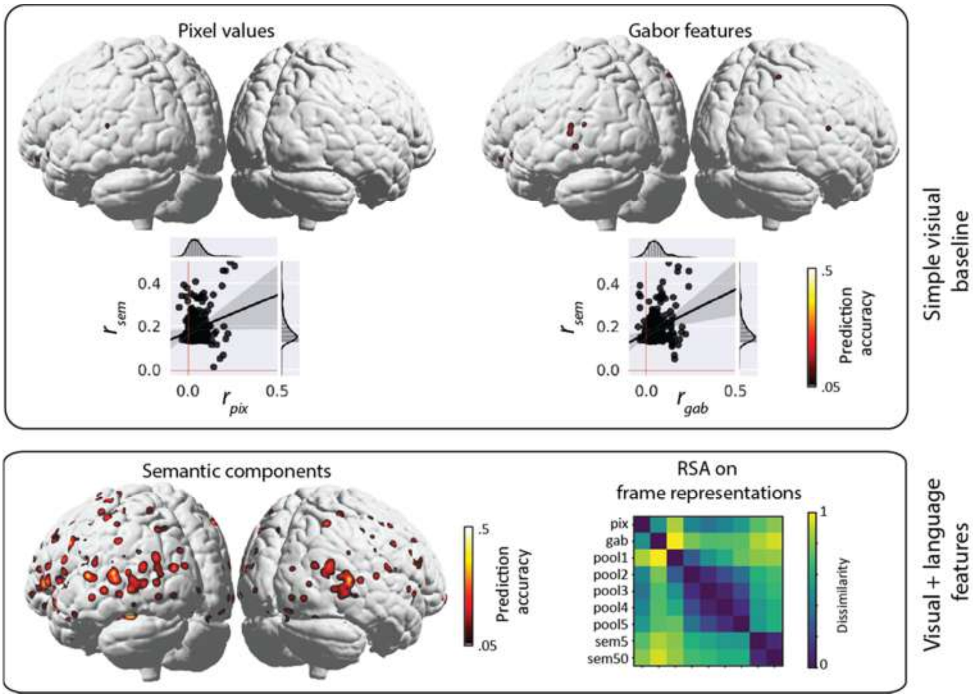
**Supplementary Figure 5.3S. Control for audio envelope and block design in the stimulus**. **(a)** Ordinary least squares fit of the HFB responses based on the block design in the film soundtrack. The fit was performed per each lag within 1 second after the audio onset to account for the neural delay during sensory processing. The displayed *t*-values correspond to the fit at the best time lag (at $p < .001$, Bonferroni corrected for the number of electrodes and number of lags). Left panel shows the cortical distribution of electrodes with a significant fit at their individual best fitted lag. Top right panel shows the distribution over the cortical labels for all the electrodes with a significant fit. Bottom right panel shows a distribution of best lags for the electrodes with a significant fit. **(b)** Cross-correlation to the audio envelope. Left panel shows best correlation within 1 second after the audio onset (at $p < .001$, Bonferroni corrected for the number of electrodes and lags). Right panel shows distribution over cortical regions and time lags.



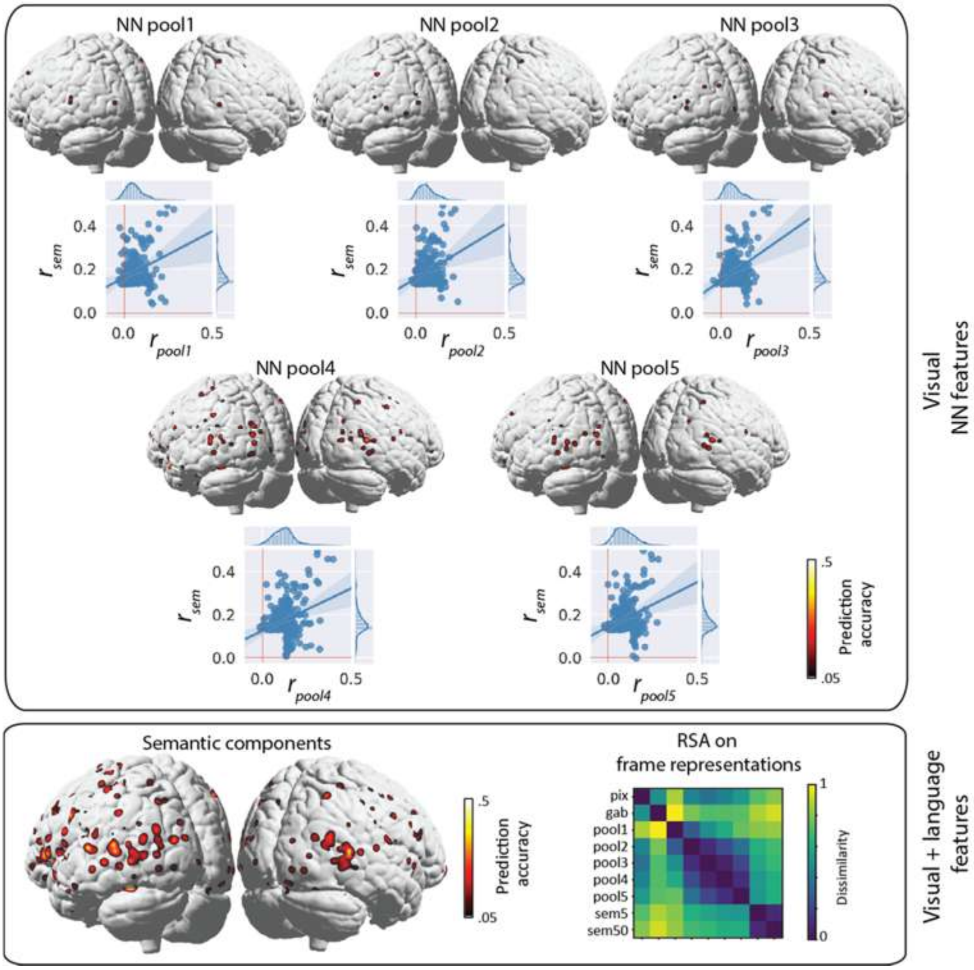
**Supplementary Figure 5.4S. Autocorrelation of semantic components.** Colored line refer to the autocorrelation of the top five semantic components (same color scheme as in Figures 2 and 6). Bold black line shown mean autocorrelation across all 50 semantic components. Thin black lines represent remaining 45 semantic components.



**Supplementary Figure 5.5S. Remaining clusters.** These are the remaining three clusters corresponding capturing distributed functional cortical networks associated with each individual semantic component. The display follows the same format as in main **Figure 5.5**.



**Supplementary Figure 5.6S. Control for low-level visual features.** Plots show comparison of the HFB fit on semantic components with a fit on low-level visual features, including colored pixel values and Gabor features modeling responses of complex cells in primary visual cortex. The ridge linear fit was performed the same way as when using the semantic components as input features. Shown are the cortical maps of the significant fit (at $p < .001$, Bonferroni corrected for the number of electrodes), scatter plots comparing prediction accuracy between each low-level visual feature set and semantic components and the dissimilarity matrix, comparing frame representations across various visual features (low-level and intermediate layers of the object recognition model) and the semantic components (only top five or all 50 components).

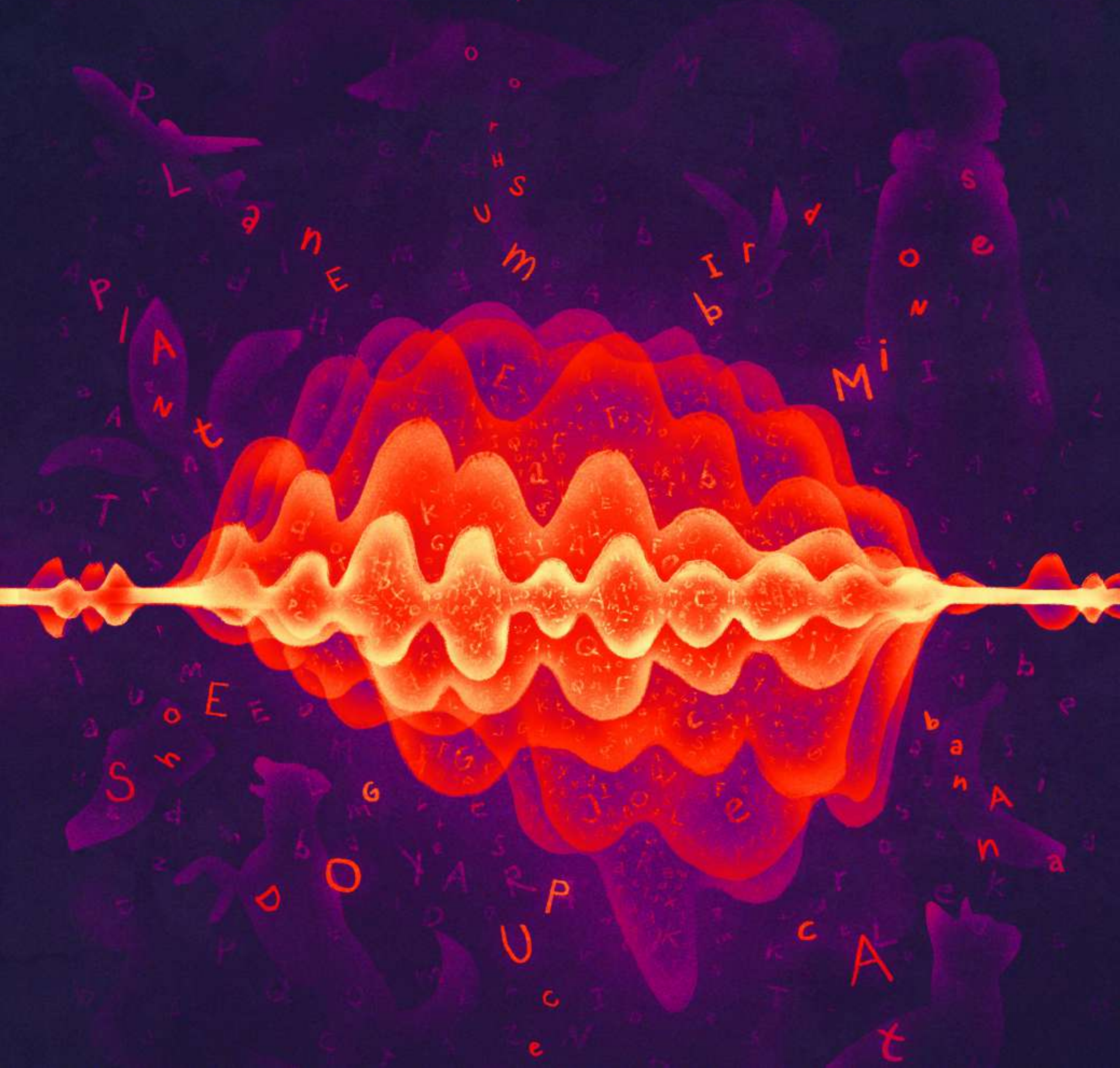


**Supplementary Figure 5.7S. Gradual emergence of semantic components from the visual features.** Plots show comparison of the HFB fit on semantic components with the fit on features of the intermediate layers of the automatic visual object recognition model. The ridge linear fit was performed the same way as when using the semantic components as input features. Shown are the cortical maps of the significant fit (at $p < .001$, Bonferroni corrected for the number of electrodes), scatter plots comparing prediction accuracy between each pooling layer feature set and semantic components and the dissimilarity matrix, comparing frame representations across various visual features (low-level and intermediate layers of the object recognition model) and the semantic components (only top five or all 50 components).

chapter

# 6

# Summary and discussion



# CHAPTER 6. SUMMARY AND DISCUSSION

## 6.1 SUMMARY

In this thesis we investigated the cortical responses during speech and language processing by applying data-driven models to ECoG data collected during perception of an audiovisual narrative. We observed that various auditory, visual and language-related properties emerged from bottom-up exploration and computational modeling of the neural data. We show that the adapted computational approaches provide data-driven support for several influential theories of language processing in the brain. At the same time, we were able to report novel findings that challenge some current theoretical views and extend our overall understanding of the cortical mechanisms underlying speech and language processing.

In **Chapter 2**, we used low-level acoustic speech features to fit the neural ECoG responses during continuous speech perception in a film-watching experiment. The fit was successful for many ECoG electrodes along the perisylvian cortex. Investigation of the tuning profiles revealed a gradient of speech information propagation along the perisylvian cortex: posterior superior temporal sites encoded fast temporal information within short time windows, whereas anterior superior temporal sites and inferior frontal gyrus encoded slow temporal information within larger time windows. In addition, we showed that the observed differences in temporal tuning along the gradient mapped well on the differences in rates of hierarchical linguistic units: phoneme rates mapped onto the fast temporal modulations and words mapped onto the slow temporal modulations. Our results confirmed and extended the previous findings regarding the increasing temporal windows of information processing in the temporal cortex (Hullett et al., 2016; Poeppel, 2003). Using a combination of a linear neural encoding model and a clustering approach, we were the first to show that these results extend to inferior frontal cortex and support the idea of integration of language-related information in this region.

In **Chapter 3**, we took a step further from the approach used in Chapter 2 and developed an entirely data-driven approach to extract non-linear auditory features from the raw sound signal in a way that best explained the associated cortical responses. For this, we trained a deep artificial neural network (ANN) on the raw sound of the full-length feature film directly to predict the ECoG brain data collected during watching of the film. Our findings showed a successful fit of our non-linear neural encoding model in all subjects. Importantly, the fit was best compared to alternative sound feature sets (including the features used in Chapter 2) and the model generalized well across subjects and experimental material when validated on data from the short film-watching experiment. In addition, we observed that the ANN architecture (a recurrent convolutional ANN) was an important factor in achieving top performance suggesting that both recurrent and convolutional mechanisms might be employed by the human brain in processing complex auditory information. The features extracted by the ANN offered data-driven support for the idea of information propagation along the perisylvian cortex put forward in Chapter 2.

In **Chapter 4**, we focused on the cortical responses to a film beyond the perisylvian cortex. In this project, we made use of high-density ECoG recordings that allowed us to investigate the functional mapping in the sensorimotor cortex at a higher degree of spatial detail than is provided with standard ECoG grids. Our main finding was identification of a distinct region in anterior dorsal precentral cortex involved in tracking of perceived speech. The identified region followed various perceptually relevant auditory properties of perceived speech, including but not limited to the spectral envelope of speech, its rhythmic phrasal pattern and pitch. Anterior dorsal precentral cortex even showed the ability to filter out background noise from the perceived speech signal. In addition, we were able to show that the identified region in precentral cortex was distinct from the adjacent hand motor region (located more posteriorly) and face motor region (located more ventrally). We discussed our findings in the context of the present theories on the involvement of (pre)motor regions (overlapping with precentral gyrus) in speech processing (Bengtsson et al., 2009; Liberman and Mattingly, 1985; Meister et al., 2007; Wilson et al., 2004).

So far, the previous chapters have focused on basic auditory features of the perceived speech signal. However, perception of speech and language cannot be reduced to processing of its low-level perceptual characteristics. One of the most important and challenging to study aspects of language is semantics and its connection with conceptual representations. Thus, in **Chapter 5**, we focused on the neural encoding of higher-level semantic features associated with perception of a visual narrative, such as a feature film. Specifically, we employed a combination of visual object recognition and language model ANNs to extract semantic information associated with different fragments of the film. The semantic features extracted by ANNs captured important semantic distinctions, such as presence or absence people in the film, presence or absence of human faces, static scene versus moments with action and movement. Importantly, these principal semantic distinctions mapped onto the neural activity within distinct distributed cortical networks. Several cortical sites associated with individual semantic dimensions were previously implicated in semantic processing consistent with our results. We also show that processing of the film features by the language model ANN improved the fit to the neural responses compared to the perceptual features extracted by the visual object recognition ANN underlying the importance of contextual and language specific information in high-level processing of perceptual input.

# 6.2 DISCUSSION

## Discussion in the context of neuroscience of language and speech

Overall, these results show that the approach taken here that combined data-driven techniques with ECoG neural recordings collected under naturalistic experimental paradigms, helps deepen our understanding of the cortical mechanisms underlying speech and language processing. To put our results in a larger context, we evaluated our findings against existing theories of the neurobiology of language and speech. Several of our findings were consistent with the dual-stream theory of language (Hickok and Poeppel, 2004) revealing evidence of propagation of the perceived speech information along the perisylvian cortex for higher-level language comprehension (Chapters 1 and 2). We also showed that the perceptual speech information was ultimately delivered to inferior frontal gyrus (Chapter 1), which aligned with the Memory, Unification and Control theory (Hagoort, 2013) underlining the integrative function of this region during speech perception. More globally, we presented evidence that perceptual processes be it in visual or auditory modality might share a common basis in that they rely on a gradient of progressively more complex and integrated information passed from early sensory cortex to higher associative areas.

At the same time, employing data-driven approaches supplied us with evidence that sparked discussion of some of the current theories related to speech perception. We expressed our concerns about some parts of these theories indicating that further work is necessary to revise them in order to incorporate and reconcile existing evidence. Specifically, the main finding of Chapter 3 led to a discussion and gives rise to revision of the current views on the involvement of (pre)motor cortex in speech perception (Keitel et al., 2018; Meister et al., 2007; Wilson et al., 2004), where we pointed out that the unifying theory about the function of the region in the light of all presented evidence is missing as of yet.

Data-driven approaches employed here also showed potential for generating new, data-driven hypotheses. More concretely, some of our findings from Chapter 2 suggested that some anticipatory prediction processes might take place in associative temporal cortex during audiovisual continuous speech perception. The results of Chapter 2 also pointed towards a complex relationship between auditory and visual sensory cortices during perception of an audiovisual material such that perception of speech might lead to a decrease of activity in high-level visual cortices possibly due to attentional shifts. The results of Chapter 4, while aligning with previous knowledge about cortical processing of visual semantics, pointed towards the possibility that semantic information processed by the human brain may not be discrete as generally considered by traditional research (Grill-Spector et al., 2001; Haxby et al., 2001; Kriegeskorte et al., 2008b) but continuous in nature. In this case, the continuous information along different semantic dimensions can be mapped onto activity of distinct cortical networks, which we showed in Chapter 4.

## Discussion of the methodological aspect of the thesis

Regarding the methodological decisions taken in this thesis, we can confirm that naturalistic paradigms provided us with rich multi-purpose datasets suitable for studying many aspects of speech processing. This choice of the experimental paradigm allowed us to investigate the neural responses to low-level perceptual auditory properties, its relationship with higher-level linguistic units (from phonemes to words), temporal rhythmic structures in speech and semantic representations of the perceived narrative. All these aspects of perceived language and speech were analyzed in neural data collected from only two film-watching experiments.

Employing naturalistic experimental paradigms also allowed us to focus on perception of continuous speech and other sensory input. We found that many of our findings verify what is known about perception of speech from experiments with more constrained language material. At the same time, using a continuous stream of speech leads to less artificial confounds in our data and improves the overall ecological validity of our results.

We also confirm that ECoG neural recording modality offers a great detail of the temporal and spatial aspects of the neural signal. We were able to minimize limitations of ECoG grids associated with sparse spatial sampling (no recordings from tissue between the electrodes) and possible cognitive deficits due to epilepsy by pooling data across a large number of subjects. Importantly, we found that high-density ECoG recordings are particularly informative and may be necessary for detailed functional mapping of the cortex involved in speech processing.

In order to relate the ECoG brain responses to the features of perceived experimental material we made heavy use of several machine learning techniques. Notably, we report that employing various optimization techniques and advanced non-linear models allowed us to achieve top accuracy in fitting the brain data on the basis of different stimulus features. This was combined with strong generalization capacity of the tested models. For instance in Chapter 2, our non-linear neural encoding model of auditory perception generalized to unseen subjects and new stimulus material, and in this new data it still provided top performance compared to the control models.

At the same time, we came across a number of challenges associated with exploration and interpretation of many results provided by our bottom-up approaches. Interpretation of the features extracted by the deep ANN in Chapter 2 was particularly challenging as unlike ANNs trained on images, ANNs trained on other data are more difficult to visualize and relate to the output labels. In general, often we relied on post-hoc analyses to relate our findings to existing language constructs or current theories of neural processing. Thus, our conclusion is that despite the temptation to focus on the best performance in explaining the cortical patterns, one should invest in the interpretability of the neural encoding model by considering the biological plausibility of its underlying computational mechanisms and incorporation of the existing experimental evidence into their model.

### Future directions and concluding remarks

We can suggest a number of future directions for our work. Expanding on the topics addressed in this thesis, we propose further investigation of the predictive mechanisms during naturalistic speech perception. Predictions occur at different levels of language hierarchy and it is interesting to study interactions of different levels using bottom-up approaches. Apart from the predictive language processing, we think that more attention should be directed at the audiovisual perception of speech to investigate whole-brain mechanisms of multimodal interactions. In particular, we are interested in testing the attentional shift hypothesis generated by the results of Chapter 2. Related to this, it would be interesting to complement the findings on the perception of visual semantics from Chapter 4 with analyses on the semantic processing of the auditory stream of the feature film.

Another promising extension of our work is inclusion of other components of the ECoG signal. Specifically, the ECoG responses in time domain could contain information related to well-known language specific evoked potentials, such as N400 and P600. Apart from that, although most of the current neural encoding and decoding work is currently focused on high-frequency band ECoG signal, low-frequency band components of ECoG and MEG signals (delta and theta range) have also been implicated in language related processing (Hermes et al., 2014; Keitel et al., 2018). A comprehensive model explaining how language induced information is transferred and combined through different bands of the neural signal is needed in the field. In addition, as we saw in Chapter 3, denser sampling of the neural tissue with high-density ECoG can be more informative about the underlying cortical function. We argue that high-density and deep intracranial recordings (stereotactic ECoG) should be combined with standard ECoG for studying whole-brain detailed neural responses to speech perception.

When it comes to the computational modeling of how speech processing takes place in the human brain, we suggest focusing more on the architecture of the employed models. We believe that the models used in this thesis can be further improved by injecting more information about the nature of the language data as well as the properties of the neural signals. Implementation of the biological and linguistic constraints should reflect in the choice of the computational mechanisms and enable important processes such as top-down and bottom-up connectivity between cortical regions, serial processing of speech with access to only past data and heavy interactions with other cognitive systems (for example memory and attention). Better neural encoding models for language and speech should also be flexible enough to accommodate various neural recording modalities and be able to incorporate behavioral evidence. We believe that some work on implementing such models has already begun (Seeliger et al., 2019), however in the present state they are optimized for other types of sensory processing. More effort is needed to adapt these models to speech and language processing and turn them into fully functional systems that model complex human brain processing.

Overall, in this work we showed that application of data-driven approaches to modeling of the neural responses to language and speech is a promising venue that can help further our understanding of the neural mechanisms of speech and language. We were able to demonstrate that the data-driven findings can be helpful in refining the current neurobiological theories of language perception. We believe that the employed methods have potential for advancing both theoretical and applied neuroscience by providing high-performing computational models of the brain responses to speech.

# BIBLIOGRAPHY

Abnar, S., Ahmed, R., Mijnheer, M., Zuidema, W., 2017. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. ArXiv Prepr. ArXiv171109285.

Andersen, A.H., Gash, D.M., Avison, M.J., 1999. Principal component analysis of the dynamic response measured by fMRI: a generalized linear systems framework. Magn. Reson. Imaging 17, 795–815.

Arai, T., Pavel, M., Hermansky, H., Avendano, C., 1999. Syllable intelligibility for temporally filtered LPC cepstral trajectories. J. Acoust. Soc. Am. 105, 2783–2791. https://doi.org/10.1121/1.426895

Arai, T., Pavel, M., Hermansky, H., Avendano, C., 1996. Intelligibility of speech with filtered time trajectories of spectral envelopes, in: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. IEEE, pp. 2490–2493.

Arnal, L.H., Giraud, A.-L., 2012. Cortical oscillations and sensory predictions. Trends Cogn. Sci. 16, 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 95–113.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95, 709-721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Barsalou, L.W., 1999. Perceptual symbol systems. Behav. Brain Sci. 22, 577–660.

Barsalou, L.W., Kyle Simmons, W., Barbey, A.K., Wilson, C.D., 2003. Grounding conceptual knowledge in modality-specific systems. Trends Cogn. Sci. 7, 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3

Bartels, A., Zeki, S., 2004. Functional brain mapping during free viewing of natural scenes. Hum. Brain Mapp. 21, 75–85. https://doi.org/10.1002/hbm.10153

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., Fries, P., 2015. Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron 85, 390–401. https://doi.org/10.1016/j.neuron.2014.12.018

Beauchamp, M.S., 2005. See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. Curr. Opin. Neurobiol. 15, 145–153.

Begliomini, C., Nelini, C., Caria, A., Grodd, W., Castiello, U., 2008. Cortical activations in humans grasp-related areas depend on hand used and handedness. PLoS ONE 3, e3388. https://doi.org/10.1371/journal.pone.0003388

Bendixen, A., SanMiguel, I., Schröger, E., 2012. Early electrophysiological indicators for predictive processing in audition: A review. Int. J. Psychophysiol. 83, 120–131. https://doi.org/10.1016/j.ijpsycho.2011.08.003

Bengtsson, S.L., Ullen, F., Ehrsson, H.H., Hashimoto, T., Kito, T., Naito, E., Forssberg, H., Sadato, N., 2009. Listening to rhythms activates motor and premotor cortices. cortex 45, 62–71.

Berezutskaya, J., Freudenburg, Z., Ramsey, N., Güçlü, U., van Gerven, M., Duivesteijn, W., Pechenizkiy, M., Fletcher, G., Menkovski, V., Postma, E., others, 2017a. Modeling brain responses to perceived speech with LSTM networks, in: Duivesteijn, W.; Pechenizkiy, M.; Fletcher, GHL (Ed.), Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning, Technische Universiteit Eindhoven, 9-10 June 2017. [Sl: sn], pp. 149–153.

Berezutskaya, J., Freudenburg, Z.V., Güçlü, U., van Gerven, M.A., Ramsey, N.F., 2017b. Neural tuning to low-level features of speech throughout the perisylvian cortex. J. Neurosci. 37, 7906–7920.

Bian, J., Gao, B., Liu, T.-Y., 2014. Knowledge-powered deep learning for word embedding, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 132–148.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528. https://doi.org/10.1093/cercor/10.5.512

Bishop, J., Keating, P., 2012. Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. J. Acoust. Soc. Am. 132, 1100–1112.

Bleichner, M.G., Jansma, J.M., Salari, E., Freudenburg, Z.V., Raemaekers, M., Ramsey, N.F., 2015. Classification of mouth movements using 7 T fMRI. J. Neural Eng. 12, 066026. https://doi.org/10.1088/1741-2560/12/6/066026

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Proceedings of the Institute of Phonetic Sciences. Amsterdam, pp. 97–110.

Boersma, P., Weenink, D., 2016. Praat: doing phonetics by computer [Computer program], Version 6.0. 14.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.

Böttger, S., Haberl, R., Prosiegel, M., Audebert, H., Rumberg, B., Forsting, M., Gizewski, E.R., 2010. Differences in cerebral activation during perception of optokinetic computer stimuli and video clips of living animals: An fMRI study. Brain Res. 1354, 132–139.

Botvinick, M.M., 2008. Hierarchical models of behavior and prefrontal function. Trends Cogn. Sci. 12, 201–208. https://doi.org/10.1016/j.tics.2008.02.009

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327.

Branco, M.P., Gaglianese, A., Glen, D.R., Hermes, D., Saad, Z.S., Petridou, N., Ramsey, N.F., 2018. ALICE: a tool for automatic localization of intra-cranial electrodes for clinical and high-density grids. J. Neurosci. Methods 301, 43–51.

Brennan, J., 2016. Naturalistic sentence comprehension in the brain. Lang. Linguist. Compass 10, 299–313.

Brennan, J., Pylkkänen, L., 2012. The time-course and spatial distribution of brain activity associated with sentence processing. NeuroImage 60, 1139–1148. https://doi.org/10.1016/j.neuroimage.2012.01.030

Broca, P., 1861. Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). Bull. Société Anat. 6, 330–57.

Brown, S., Laird, A.R., Pfordresher, P.Q., Thelen, S.M., Turkeltaub, P., Liotti, M., 2009. The somatotopy of speech: phonation and articulation in the human motor cortex. Brain Cogn. 70, 31–41.

Brown, S., Ngan, E., Liotti, M., 2007. A larynx area in the human motor cortex. Cereb. Cortex 18, 837–845.

Brugman, H., Russel, A., Nijmegen, X., 2004. Annotating multi-media/multi-modal resources with ELAN, in: LREC.

Cadieu, C.F., Hong, H., Yamins, D.L., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J., 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput. Biol. 10, e1003963.

Callan, D.E., Jones, J.A., Callan, A.M., Akahane-Yamada, R., 2004. Phonetic perceptual identification by native-and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. NeuroImage 22, 1182–1194.

Callan, D.E., Jones, J.A., Munhall, K., Callan, A.M., Kroos, C., Vatikiotis-Bateson, E., 2003. Neural processes underlying perceptual enhancement by visual speech gestures. Neuroreport 14, 2213–2218.

Carey, M.J., Parris, E.S., Lloyd-Thomas, H., Bennett, S., 1996. Robust prosodic features for speaker identification, in: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. IEEE, pp. 1800–1803.

Carota, F., Kriegeskorte, N., Nili, H., Pulvermüller, F., 2017. Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. Cereb. Cortex cercor;bhw379v1. https://doi.org/10.1093/cercor/bhw379

Cate, A.D., Herron, T.J., Kang, X., Yund, E.W., Woods, D.L., 2012. Intermodal attention modulates visual processing in dorsal and ventral streams. NeuroImage 63, 1295–1304. https://doi.org/10.1016/j.neuroimage.2012.08.026

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., Knight, R.T., 2010. Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. 13, 1428–1432. https://doi.org/10.1038/nn.2641

Chao, L.L., Haxby, J.V., Martin, A., 1999. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat. Neurosci. 2, 913.

Chartier, J., Anumanchipalli, G.K., Johnson, K., Chang, E.F., 2018. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron 98, 1042-1054.e4. https://doi.org/10.1016/j.neuron.2018.04.031

Chen, J.L., Zatorre, R.J., Penhune, V.B., 2006. Interactions between auditory and dorsal premotor cortex during synchronization to musical rhythms. Neuroimage 32, 1771–1781.

Chen, Y., Davis, M.H., Pulvermüller, F., Hauk, O., 2013. Task modulation of brain responses in visual word recognition as studied using EEG/MEG and fMRI. Front. Hum. Neurosci. 7, 376.

Cheung, C., Hamilton, L.S., Johnson, K., Chang, E.F., 2016. The auditory representation of speech sounds in human motor cortex. Elife 5, e12577.

Chi, T., Ru, P., Shamma, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. 118, 887. https://doi.org/10.1121/1.1945807

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. 6, 27755.

Cichy, R.M., Pantazis, D., Oliva, A., 2014. Resolving human object recognition in space and time. Nat. Neurosci. 17, 455.

Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., Pesaran, B., 2014. Sensory–motor transformations for speech occur bilaterally. Nature 507, 94.

Collier, R., 1975. Physiological correlates of intonation patterns. J. Acoust. Soc. Am. 58, 249–255.

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618.

Crinion, J.T., Lambon-Ralph, M.A., Warburton, E.A., Howard, D., Wise, R.J.S., 2003. Temporal lobe regions engaged during normal speech comprehension. Brain 126, 1193–1201. https://doi.org/10.1093/brain/awg104

Crone, Ne., Hao, L., Hart, J., Boatman, D., Lesser, R.P., Irizarry, R., Gordon, B., 2001. Electrocorticographic gamma activity during word production in spoken and sign language. Neurology 57, 2045–2053.

Crone, N.E., Miglioretti, D.L., Gordon, B., Lesser, R.P., 1998. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain J. Neurol. 121, 2301–2315.

Dalal, S.S., Baillet, S., Adam, C., Ducorps, A., Schwartz, D., Jerbi, K., Bertrand, O., Garnero, L., Martinerie, J., Lachaux, J.-P., 2009. Simultaneous MEG and intracranial EEG recordings during attentive reading. NeuroImage 45, 1289–1304. https://doi.org/10.1016/j.neuroimage.2009.01.017

D'Ausilio, A., Craighero, L., Fadiga, L., 2012. The contribution of the frontal lobe to the perception of speech. J. Neurolinguistics 25, 328–335.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., Fadiga, L., 2009. The motor somatotopy of speech perception. Curr. Biol. 19, 381–385. https://doi.org/10.1016/j.cub.2009.01.017

de Gelder, B., De Borst, A., Watson, R., 2015. The perception of emotion in body expressions. Wiley Interdiscip. Rev. Cogn. Sci. 6, 149–158.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37, 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

Dehaene-Lambertz, G., Dupoux, E., Gout, A., 2000. Electrophysiological correlates of phonological processing: a cross-linguistic study. J. Cogn. Neurosci. 12, 635–647. https://doi.org/10.1162/089892900562390

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv Prepr. ArXiv181004805.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti, G., 1992. Understanding motor events: a neurophysiological study. Exp. Brain Res. 91, 176–180.

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73, 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Dichter, B.K., Breshears, J.D., Leonard, M.K., Chang, E.F., 2018. The control of vocal pitch in human laryngeal motor cortex. Cell 174, 21–31.

Dilks, D.D., Julian, J.B., Paunov, A.M., Kanwisher, N., 2013. The occipital place area is causally and selectively involved in scene perception. J. Neurosci. 33, 1331–1336.

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158.

Ding, N., Simon, J.Z., 2012. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107, 78–89. https://doi.org/10.1152/jn.00297.2011

Donders, F.C., 1969. On the speed of mental processes. Acta Psychol. (Amst.) 30, 412–431.

Dronkers, N.F., Wilkins, D.P., Van Valin Jr, R.D., Redfern, B.B., Jaeger, J.J., 2004. Lesion analysis of the brain areas involved in language comprehension. Cognition 92, 145–177.

Drullman, R., 1995. Temporal envelope and fine structure cues for speech intelligibility. J. Acoust. Soc. Am. 97, 585–592.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95, 2670–2680. https://doi.org/10.1121/1.409836

Du, Y., Buchsbaum, B.R., Grady, C.L., Alain, C., 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. 111, 7126–7131. https://doi.org/10.1073/pnas.1318738111

Duffau, H., Capelle, L., Denvil, D., Gatignol, P., Sichez, N., Lopes, M., Sichez, J.-P., Van Effenterre, R., 2003. The role of dominant premotor cortex in language: a study using intraoperative functional mapping in awake patients. Neuroimage 20, 1903–1914.

Dushanova, J., Donoghue, J., 2010. Neurons in primary motor cortex engaged during action observation. Eur. J. Neurosci. 31, 386–398. https://doi.org/10.1111/j.1460-9568.2009.07067.x

Elliott, T.M., Theunissen, F.E., 2009. The modulation transfer function for speech intelligibility. PLoS Comput Biol 5, e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598.

Fadiga, L., Craighero, L., Olivier, E., 2005. Human motor cortex excitability during the perception of others' action. Curr. Opin. Neurobiol. 15, 213–218. https://doi.org/10.1016/j.conb.2005.03.013

Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Flinker, A., Chang, E., Barbaro, N., Berger, M., Knight, R., 2011. Sub-centimeter language organization in the human temporal lobe. Brain Lang. 117, 103–109.

Flinker, A., Korzeniewska, A., Shestyuk, A.Y., Franaszczuk, P.J., Dronkers, N.F., Knight, R.T., Crone, N.E., 2015. Redefining the role of Broca's area in speech. Proc. Natl. Acad. Sci. 112, 2871–2875. https://doi.org/10.1073/pnas.1414491112

Floel, A., Ellger, T., Breitenstein, C., Knecht, S., 2003. Language perception activates the hand motor cortex: implications for motor theories of speech perception. Eur. J. Neurosci. 18, 704–708. https://doi.org/10.1046/j.1460-9568.2003.02774.x

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., Giraud, A.-L., 2014. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nat. Commun. 5, 4694. https://doi.org/10.1038/ncomms5694

Formisano, E., Martino, F.D., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322, 970–973. https://doi.org/10.1126/science.1164318

Foti, D., Roberts, F., 2016. The neural dynamics of speech perception: Dissociable networks for processing linguistic content and monitoring speaker turn-taking. Brain Lang. 157, 63–71.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. Science 315, 972–976.

Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends Cogn. Sci. 16, 262–268. https://doi.org/10.1016/j.tics.2012.04.001

Friederici, A.D., 2011. The brain basis of language processing: from structure to function. Physiol. Rev. 91, 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc. B Biol. Sci. 360, 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K., Frith, C., Liddle, P., Frackowiak, R., 1993. Functional connectivity: the principal-component analysis of large (PET) data sets. J. Cereb. Blood Flow Metab. 13, 5–14.

Fu, Q.-J., Shannon, R.V., 2000. Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners. J. Acoust. Soc. Am. 107, 589–597. https://doi.org/10.1121/1.428325

Gaglianese, A., Harvey, B.M., Vansteensel, M.J., Dumoulin, S.O., Ramsey, N.F., Petridou, N., 2017. Separate spatial and temporal frequency tuning to visual motion in human MT+ measured with ECoG. Hum. Brain Mapp. 38, 293–307.

Galantucci, B., Fowler, C.A., Turvey, M.T., 2006. The motor theory of speech perception reviewed. Psychon. Bull. Rev. 13, 361–377.

Ganong, W.F., 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Hum. Percept. Perform. 6, 110.

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., Kleinschmidt, A., 2000. Representation of the temporal envelope of sounds in the human brain. J. Neurophysiol. 84, 1588–1598.

Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517. https://doi.org/10.1038/nn.3063

Glanz, O., Derix, J., Kaur, R., Schulze-Bonhage, A., Auer, P., Aertsen, A., Ball, T., 2018. Real-life speech production and perception have a shared premotor-cortical substrate. Sci. Rep. 8. https://doi.org/10.1038/s41598-018-26801-x

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 249–256.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends Neurosci. 15, 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Grafton, S.T., de C. Hamilton, A.F., 2007. Evidence for a distributed hierarchy of action representation in the brain. Hum. Mov. Sci. 26, 590–616. https://doi.org/10.1016/j.humov.2007.05.009

Greenfield, P.M., 1991. Language, tools and brain: the ontogeny and phylogeny of hierarchically organized sequential behavior. Behav. Brain Sci. 14, 531–551.

Grill-Spector, K., Kourtzi, Z., Kanwisher, N., 2001. The lateral occipital complex and its role in object recognition. Vision Res. 41, 1409–1422. https://doi.org/10.1016/S0042-6989(01)00073-6

Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annu. Rev. Neurosci. 27, 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Grill-Spector, K., Weiner, K.S., 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci. 15, 536–548.

Groen, I.I., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., Baker, C.I., 2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7, e32962.

Grosbras, M.-H., Beaton, S., Eickhoff, S.B., 2012. Brain regions involved in human movement perception: A quantitative voxel-based meta-analysis. Hum. Brain Mapp. 33, 431–454.

Güçlü, U., Thielen, J., Hanke, M., van Gerven, M., 2016. Brains on beats, in: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 2101–2109.

Güçlü, U., van Gerven, M.A., 2017. Modeling the dynamics of human brain activity with recurrent neural networks. Front. Comput. Neurosci. 11, 7.

Güçlü, U., van Gerven, M.A., 2015a. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.

Güçlü, U., van Gerven, M.A.J., 2015b. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage. https://doi.org/10.1016/j.neuroimage.2015.12.036

Güçlü, U., van Gerven, M.A.J., 2014. Unsupervised feature learning improves prediction of human brain activity in response to natural images. PLoS Comput. Biol. 10, e1003724. https://doi.org/10.1371/journal.pcbi.1003724

Hagoort, P., 2013. MUC (memory, unification, control) and beyond. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00416

Hagoort, P., 2005. On Broca, brain, and binding: a new framework. Trends Cogn. Sci. 9, 416–423. https://doi.org/10.1016/j.tics.2005.07.004

Halai, A.D., Woollams, A.M., Ralph, M.A.L., 2017. Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: revealing the unique neural correlates of speech fluency, phonology and semantics. Cortex 86, 275–289.

Halgren, E., Dhond, R.P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J.D., Dale, A.M., 2002. N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. NeuroImage 17, 1101–1116. https://doi.org/10.1006/nimg.2002.1268

Hamilton, L.S., Huth, A.G., 2018. The revolution will not be controlled: natural stimuli in speech neuroscience. Lang. Cogn. Neurosci. 1–10. https://doi.org/10.1080/23273798.2018.1499946

Hasson, U., Chen, J., Honey, C.J., 2015. Hierarchical process memory: memory as an integral component of information processing. Trends Cogn. Sci. 19, 304–313. https://doi.org/10.1016/j.tics.2015.04.006

Hasson, U., Harel, M., Levy, I., Malach, R., 2003. Large-scale mirror-symmetry organization of human occipito-temporal object areas. Neuron 37, 1027–1041. https://doi.org/10.1016/S0896-6273(03)00144-2

Hasson, U., Malach, R., Heeger, D.J., 2010. Reliability of cortical activity during natural stimulation. Trends Cogn. Sci. 14, 40–48. https://doi.org/10.1016/j.tics.2009.10.011

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640. https://doi.org/10.1126/science.1089506

Haufe, S., DeGuzman, P., Henin, S., Arcaro, M., Honey, C.J., Hasson, U., Parra, L.C., 2018. Elucidating relations between fMRI, ECoG, and EEG through a common natural stimulus. NeuroImage 179, 79–91. https://doi.org/10.1016/j.neuroimage.2018.06.016

Hauk, O., Johnsrude, I., Pulvermüller, F., 2004. Somatotopic representation of action words in human motor and premotor cortex. Neuron 41, 301–307.

Haxby, J.V., Gobbini, M.I., 2011. Distributed neural systems for face perception. The Oxford Handbook of Face Perception.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430. https://doi.org/10.1126/science.1063736

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Heldner, M., 2011. Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. J. Acoust. Soc. Am. 130, 508–513. https://doi.org/10.1121/1.3598457

Hermes, D., Miller, K.J., Noordmans, H.J., Vansteensel, M.J., Ramsey, N.F., 2010. Automated electrocorticographic electrode localization on individually rendered brain surfaces. J. Neurosci. Methods 185, 293–298.

Hermes, D., Miller, K.J., Vansteensel, M.J., Aarnoutse, E.J., Leijten, F.S.S., Ramsey, N.F., 2012. Neurophysiologic correlates of fMRI in human motor cortex. Hum. Brain Mapp. 33, 1689–1699. https://doi.org/10.1002/hbm.21314

Hermes, D., Miller, K.J., Vansteensel, M.J., Edwards, E., Ferrier, C.H., Bleichner, M.G., van Rijen, P.C., Aarnoutse, E.J., Ramsey, N.F., 2014. Cortical theta wanes for language. NeuroImage 85 Pt 2, 738–748. https://doi.org/10.1016/j.neuroimage.2013.07.029

Hevner, K., 1937. The affective value of pitch and tempo in music. Am. J. Psychol. 49, 621–630.

Hickok, G., 2012. Computational neuroanatomy of speech production. Nat. Rev. Neurosci. 13, 135–145. https://doi.org/10.1038/nrn3158

Hickok, G., 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. J. Cogn. Neurosci. 21, 1229–1243.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402. https://doi.org/10.1038/nrn2113

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition, Towards a New Functional Anatomy of Language 92, 67–99. https://doi.org/10.1016/j.cognition.2003.10.011

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., Theunissen, F.E., 2017. Encoding and decoding models in cognitive electrophysiology. Front. Syst. Neurosci. 11. https://doi.org/10.3389/fnsys.2017.00061

Honey, C. J., Thesen, T., Donner, T.H., Silbert, L.J., Carlson, C.E., Devinsky, O., Doyle, W.K., Rubin, N., Heeger, D.J., Hasson, U., 2012. Slow cortical dynamics and the accumulation of information over long timescales. Neuron 76, 423–434. https://doi.org/10.1016/j.neuron.2012.08.011

Honey, C. J., Thompson, C.R., Lerner, Y., Hasson, U., 2012. Not lost in translation: neural responses shared across languages. J. Neurosci. 32, 15277–15283. https://doi.org/10.1523/JNEUROSCI.1800-12.2012

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Netw. 4, 251–257.

Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. 77, 1069–1077. https://doi.org/10.1121/1.392224

Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., Chang, E.F., 2016. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. J. Neurosci. 36, 2014–2026. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. https://doi.org/10.1038/nature17637

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224.

Indefrey, P., Levelt, W.J.M., 2004. The spatial and temporal signatures of word production components. Cognition 92, 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ArXiv Prepr. ArXiv150203167.

Jain, S., Huth, A., 2018. Incorporating context into language encoding models for fMRI, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 6628–6637.

Jerbi, K., Ossandón, T., Hamamé, C.M., Senova, S., Dalal, S.S., Jung, J., Minotti, L., Bertrand, O., Berthoz, A., Kahane, P., Lachaux, J.-P., 2009. Task-related gamma-band dynamics from an intracerebral perspective: Review and implications for surface EEG and MEG. Hum. Brain Mapp. 30, 1758–1771. https://doi.org/10.1002/hbm.20750

Jones, E., Oliphant, T., Peterson, P., others, 2001. SciPy: Open source scientific tools for Python.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Karpathy, A., Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3128–3137.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352–355.

Keitel, A., Gross, J., Kayser, C., 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLOS Biol. 16, e2004473. https://doi.org/10.1371/journal.pbio.2004473

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H., 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98, 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Kendall, M.G., Stuart, A., 1973. The advanced theory of statistics. Vol. 2: inference and relationship.

Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Comput. Biol. 10, e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kiani, R., Esteky, H., Mirpour, K., Tanaka, K., 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J. Neurophysiol. 97, 4296–4309.

Kietzmann, T.C., McClure, P., Kriegeskorte, N., 2018. Deep neural networks in computational neuroscience. bioRxiv 133504.

Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K., Cichy, R.M., Hauk, O., Kriegeskorte, N., 2019. Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. 116, 21854–21863.

Kitamura, T., Akagi, M., 1995. Speaker individualities in speech spectral envelopes. J. Acoust. Soc. Jpn. E 16, 283–289.

Kravitz, D.J., Saleem, K.S., Baker, C.I., Mishkin, M., 2011. A new neural framework for visuospatial processing. Nat. Rev. Neurosci. 12, 217–230. https://doi.org/10.1038/nrn3008

Krieger-Redwood, K., Jefferies, E., 2014. TMS interferes with lexical-semantic retrieval in left inferior frontal gyrus and posterior middle temporal gyrus: Evidence from cyclical picture naming. Neuropsychologia 64, 24–32.

Kriegeskorte, N., 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. 1, 417–446.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008a. Representational similarity analysis–connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems. pp. 1097–1105.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., Schalk, G., 2013. The tracking of speech envelope in the human cortex. PLoS ONE 8, e53398. https://doi.org/10.1371/journal.pone.0053398

Lachaux, J.-P., Fonlupt, P., Kahane, P., Minotti, L., Hoffmann, D., Bertrand, O., Baciu, M., 2007. Relationship between task-related gamma oscillations and BOLD signal: new insights from combined fMRI and intracranial EEG. Hum. Brain Mapp. 28, 1368–1375. https://doi.org/10.1002/hbm.20352

Lahnakoski, J.M., Glerean, E., Salmi, J., Jääskeläinen, I.P., Sams, M., Hari, R., Nummenmaa, L., 2012. Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Front. Hum. Neurosci. 6, 233.

Leopold, D.A., Rhodes, G., 2010. A comparative view of face perception. J. Comp. Psychol. 124, 233.

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31, 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychol. Rev. 74, 431.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Lin, H.W., Tegmark, M., Rolnick, D., 2017. Why does deep and cheap learning work so well? J. Stat. Phys. 168, 1223–1247.

Loeb, G.E., Brown, I.E., Cheng, E.J., 1999. A hierarchical foundation for models of sensorimotor control. Exp. Brain Res. 126, 1–18.

Maaten, L. van der, Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Martin, A., 2007. The representation of object concepts in the brain. Annu. Rev. Psychol. 58, 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N.E., Rieger, J., Schalk, G., Knight, R.T., Pasley, B.N., 2014. Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroengineering 7. https://doi.org/10.3389/fneng.2014.00014

Martin, S., Mikutta, C., Leonard, M.K., Hungate, D., Koelsch, S., Chang, E.F., Millan, J. del R., Knight, R.T., Pasley, B.N., 2017. Neural encoding of auditory features during music perception and imagery: insight into the brain of a piano player. bioRxiv. https://doi.org/10.1101/106617

Mayka, M.A., Corcos, D.M., Leurgans, S.E., Vaillancourt, D.E., 2006. Three-dimensional locations and boundaries of motor and premotor cortices as defined by functional brain imaging: a meta-analysis. Neuroimage 31, 1453–1474.

Mazaika, P.K., Hoeft, F., Glover, G.H., Reiss, A.L., 2009. Methods and software for fMRI analysis of clinical subjects. Neuroimage 47, S58.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748. https://doi.org/10.1038/264746a0

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., Iacoboni, M., 2007. The essential role of premotor cortex in speech perception. Curr. Biol. 17, 1692–1696.

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

Mhaskar, H., Liao, Q., Poggio, T., 2016. Learning functions: when is deep better than shallow. ArXiv Prepr. ArXiv160300988.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. ArXiv Prepr. ArXiv13013781.

Mishkin, M., Ungerleider, L.G., Macko, K.A., 1983. Object vision and spatial vision: two cortical pathways. Trends Neurosci. 6, 414–417.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. science 320, 1191–1195.

Morillon, B., Schroeder, C.E., Wyart, V., 2014. Motor contributions to the temporal precision of auditory attention. Nat. Commun. 5, 5255.

Möttönen, R., Watkins, K.E., 2012. Using TMS to study the role of the articulatory motor system in speech perception. Aphasiology 26, 1103–1118. https://doi.org/10.1080/02687038.2011.619515

Moutoussis, K., Zeki, S., 1997. Functional segregation and temporal hierarchy of the visual perceptive systems. Proc. R. Soc. Lond. B Biol. Sci. 264, 1407–1414.

Murakami, T., Restle, J., Ziemann, U., 2011. Observation-execution matching and action inhibition in human primary motor cortex during viewing of speech-related lip movements or listening to speech. Neuropsychologia 49, 2045–2054. https://doi.org/10.1016/j.neuropsychologia.2011.03.034

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Müsch, K., Himberger, K., Tan, K.M., Valiante, T.A., Honey, C.J., 2018. Transformation of speech sequences in human sensorimotor circuits. bioRxiv 419358.

Nakai, T., Koide-Majima, N., Nishimoto, S., 2018. Encoding and decoding of music-genre representations in the human brain, in: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Presented at the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, Miyazaki, Japan, pp. 584–589. https://doi.org/10.1109/SMC.2018.00108

Naselaris, T., Stansbury, D.E., Gallant, J.L., 2012. Cortical representation of animate and inanimate objects in complex natural scenes. J. Physiol.-Paris 106, 239–249.

Neggers, S.F., Hermans, E.J., Ramsey, N.F., 2008. Enhanced sensitivity with fast three-dimensional blood-oxygen-level-dependent functional MRI: comparison of SENSE–PRESTO and 2D-EPI at 3 T. NMR Biomed. 21, 663–676.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10, e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Oliphant, T.E., 2006. A guide to NumPy. Trelgol Publishing USA.

Olthoff, A., Baudewig, J., Kruse, E., Dechent, P., 2008. Cortical sensorimotor control in vocalization: a functional magnetic resonance imaging study. The Laryngoscope 118, 2091–2096.

O'toole, A.J., Jiang, F., Abdi, H., Haxby, J.V., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. Cogn. Neurosci. J. Of 17, 580–590.

Overath, T., Kumar, S., von Kriegstein, K., Griffiths, T.D., 2008. Encoding of spectral correlation over time in auditory cortex. J. Neurosci. 28, 13268–13273. https://doi.org/10.1523/JNEUROSCI.4596-08.2008

Partridge, M., Calvo, R.A., 1998. Fast dimensionality reduction and simple PCA. Intell. Data Anal. 2, 203–214.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10, e1001251. https://doi.org/10.1371/journal.pbio.1001251

Patterson, R., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1987. An efficient auditory filterbank based on the gammatone function, in: A Meeting of the IOC Speech Group on Auditory Modelling at RSRE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S.J., Kanwisher, N., Botvinick, M., Fedorenko, E., 2018. Toward a universal decoder of linguistic meaning from brain activation. Nat. Commun. 9, 963.

Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' Speech Commun. 41, 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Pulvermüller, F., Huss, M., Kherif, F., Martin, F.M. del P., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. 103, 7865–7870. https://doi.org/10.1073/pnas.0509989103

Raposo, A., Moss, H.E., Stamatakis, E.A., Tyler, L.K., 2009. Modulation of motor and premotor cortices by actions, action words and action sentences. Neuropsychologia 47, 388–396. https://doi.org/10.1016/j.neuropsychologia.2008.09.017

Rasmussen, C.E., 2003. Gaussian processes in machine learning, in: Summer School on Machine Learning. Springer, pp. 63–71.

Rauschecker, J.P., 1998. Cortical processing of complex sounds. Curr. Opin. Neurobiol. 8, 516–521. https://doi.org/10.1016/S0959-4388(98)80040-8

Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc. Natl. Acad. Sci. 97, 11800–11806. https://doi.org/10.1073/pnas.97.22.11800

Ray, S., Crone, N.E., Niebur, E., Franaszczuk, P.J., Hsiao, S.S., 2008. Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. J. Neurosci. 28, 11526–11536.

Ritaccio, A., Boatman-Reich, D., Brunner, P., Cervenka, M.C., Cole, A.J., Crone, N., Duckrow, R., Korzeniewska, A., Litt, B., Miller, K.J., others, 2011. Proceedings of the second international workshop on advances in electrocorticography. Epilepsy Behav. 22, 641–650.

Ritaccio, A., Brunner, P., Cervenka, M.C., Crone, N., Guger, C., Leuthardt, E., Oostenveld, R., Stacey, W., Schalk, G., 2010. Proceedings of the first international workshop on advances in electrocorticography. Epilepsy Behav. 19, 204–215.

Ritaccio, A., Brunner, P., Crone, N.E., Gunduz, A., Hirsch, L.J., Kanwisher, N., Litt, B., Miller, K., Moran, D., Parvizi, J., others, 2013. Proceedings of the fourth international workshop on advances in electrocorticography. Epilepsy Behav. 29, 259–268.

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annu Rev Neurosci 27, 169–192.

Rizzolatti, G., Luppino, G., 2001. The cortical motor system. Neuron 31, 889–901.

Rogalsky, C., Hickok, G., 2009. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. Cereb. Cortex 19, 786–796. https://doi.org/10.1093/cercor/bhn126

Roland, P.E., Larsen, B., Lassen, N.A., Skinhoj, E., 1980. Supplementary motor area and other cortical areas in organization of voluntary movements in man. J. Neurophysiol. 43, 118–136.

Roth, B.J., Hallett, M., 1992. Optimal focal transcranial magnetic activation of the human motor cortex: effects of coil orientation, shape of the induced current pulse, and stimulus intensity. J. Clin. Neurophysiol. 9, 132–136.

Salo, V.C., Ferrari, P.F., Fox, N.A., 2019. The role of the motor system in action understanding and communication: Evidence from human infants and non-human primates. Dev. Psychobiol. 61, 390–401.

SanMiguel, I., Widmann, A., Bendixen, A., Trujillo-Barreto, N., Schroger, E., 2013. Hearing silences: human auditory processing relies on preactivation of sound-specific brain activity patterns. J. Neurosci. 33, 8633–8639. https://doi.org/10.1523/JNEUROSCI.5821-12.2013

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., Formisano, E., 2014. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. 10, e1003412. https://doi.org/10.1371/journal.pcbi.1003412

Sato, M., Tremblay, P., Gracco, V.L., 2009. A mediating role of the premotor cortex in phoneme segmentation. Brain Lang. 111, 1–7.

Saur, D., Kreher, B.W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., Weiller, C., 2008. Ventral and dorsal pathways for language. Proc. Natl. Acad. Sci. 105, 18035–18040. https://doi.org/10.1073/pnas.0805234105

Schellekens, W., Petridou, N., Ramsey, N.F., 2018. Detailed somatotopy in primary motor and somatosensory cortex revealed by Gaussian population receptive fields. NeuroImage 179, 337–347. https://doi.org/10.1016/j.neuroimage.2018.06.062

Schmidt, C.F., Zaehle, T., Meyer, M., Geiser, E., Boesiger, P., Jancke, L., 2008. Silent and continuous fMRI scanning differentially modulate activation in an auditory language comprehension task. Hum. Brain Mapp. 29, 46–56.

Schönwiesner, M., Zatorre, R.J., 2009. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. U. S. A. 106, 14611–14616. https://doi.org/10.1073/pnas.0907682106

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

Scott, S.K., McGettigan, C., Eisner, F., 2009. A little more conversation, a little less action—candidate roles for the motor cortex in speech perception. Nat. Rev. Neurosci. 10, 295.

Seabold, S., Perktold, J., 2010. Statsmodels: econometric and statistical modeling with python, in: Proceedings of the 9th Python in Science Conference. Scipy, p. 61.

Seeliger, K., Ambrogioni, L., Güçlütürk, Y., Güçlü, U., van Gerven, M.A., 2019. Neural System Identification with Neural Information Flow. bioRxiv 553255.

Shtyrov, Y., Butorina, A., Nikolaeva, A., Stroganova, T., 2014. Automatic ultrarapid activation and inhibition of cortical motor systems in spoken word comprehension. Proc. Natl. Acad. Sci. 111, E1918–E1923. https://doi.org/10.1073/pnas.1323158111

Siero, J.C.W., Hermes, D., Hoogduin, H., Luijten, P.R., Ramsey, N.F., Petridou, N., 2014. BOLD matches neuronal activity at the mm scale: a combined 7T fMRI and ECoG study in human sensorimotor cortex. NeuroImage 101, 177–184. https://doi.org/10.1016/j.neuroimage.2014.07.002

Simonyan, K., Horwitz, B., 2011. Laryngeal motor cortex and control of speech in humans. The Neuroscientist 17, 197–208.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. ArXiv Prepr. ArXiv14091556.

Skipper, J.I., Devlin, J.T., Lametti, D.R., 2017. The hearing ear is always found close to the speaking tongue : review of the role of the motor system in speech perception. Brain Lang. 164, 77–105. https://doi.org/10.1016/j.bandl.2016.10.004

Skipper, J.I., Nusbaum, H.C., Small, S.L., 2006. Lending a helping hand to hearing: another motor theory of speech perception. Action Lang. Mirror Neuron Syst. 250–285.

Skipper, J.I., Nusbaum, H.C., Small, S.L., 2005. Listening to talking faces: motor cortical activation during speech perception. NeuroImage 25, 76–89. https://doi.org/10.1016/j.neuroimage.2004.11.006

Skipper, J.I., van Wassenhove, V., Nusbaum, H.C., Small, S.L., 2007. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. Cereb. Cortex 17, 2387–2399. https://doi.org/10.1093/cercor/bhl147

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C., 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Netw. 32, 323–332.

Stanley, D.A., Rubin, N., 2003. fMRI activation in response to illusory contours and salient regions in the human lateral occipital complex. Neuron 37, 323–331.

Szegedy, C., Toshev, A., Erhan, D., 2013. Deep neural networks for object detection, in: Advances in Neural Information Processing Systems. pp. 2553–2561.

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1701–1708.

Ter Keurs, M., Festen, J.M., Plomp, R., 1993. Effect of spectral envelope smearing on speech reception. II. J. Acoust. Soc. Am. 93, 1547–1552.

Ter Keurs, M., Festen, J.M., Plomp, R., 1992. Effect of spectral envelope smearing on speech reception. I. J. Acoust. Soc. Am. 91, 2872–2880.

Thielscher, A., Kammer, T., 2002. Linking physics with physiology in TMS: a sphere field model to determine the cortical stimulation site in TMS. Neuroimage 17, 1117–1130.

Tokui, S., Oono, K., Hido, S., Clayton, J., 2015. Chainer: a next-generation open source framework for deep learning, in: Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS).

Tootell, R.B., Reppas, J.B., Kwong, K.K., Malach, R., Born, R.T., Brady, T.J., Rosen, B.R., Belliveau, J.W., 1995. Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. J. Neurosci. 15, 3215–3230.

Towle, V.L., Yoon, H.A., Castelle, M., Edgar, J.C., Biassou, N.M., Frim, D.M., Spire, J.P., Kohrman, M.H., 2008. ECoG gamma activity during a language task: Differentiating expressive and receptive speech areas. Brain 131, 2013–2027. https://doi.org/10.1093/brain/awn147

Tramo, M.J., Cariani, P.A., Koh, C.K., Makris, N., Braida, L.D., 2005. Neurophysiology and neuroanatomy of pitch perception: auditory cortex. Ann. N. Y. Acad. Sci. 1060, 148–174.

Ungerleider, L.G., 1982. Two cortical visual systems. Anal. Vis. Behav. 549–586.

Van Essen, D.C., Anderson, C.H., Felleman, D.J., 1992. Information processing in the primate visual system: an integrated systems perspective. Science 255, 419–423.

Watkins, K.E., Strafella, A.P., Paus, T., 2003. Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia 41, 989–994. https://doi.org/10.1016/S0028-3932(02)00316-0

Wen, H., Shi, J., Chen, W., Liu, Z., 2018. Transferring and generalizing deep-learning-based neural encoding models across subjects. NeuroImage 176, 152–163.

Wernicke, C., 1874. Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis. Cohn.

Wheaton, K.J., Thompson, J.C., Syngeniotis, A., Abbott, D.F., Puce, A., 2004. Viewing the motion of human body parts activates different regions of premotor, temporal, and parietal cortex. Neuroimage 22, 277–288.

Wilson, S.M., Iacoboni, M., 2006. Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. NeuroImage 33, 316–325. https://doi.org/10.1016/j.neuroimage.2006.05.032

Wilson, S.M., Molnar-Szakacs, I., Iacoboni, M., 2008. Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. Cereb. Cortex 18, 230–242. https://doi.org/10.1093/cercor/bhm049

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7, 701–702. https://doi.org/10.1038/nn1263

Yamins, D.L., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356.

Yousry, T., Schmid, U., Alkadhi, H., Schmidt, D., Peraud, A., Buettner, A., Winkler, P., 1997. Localization of the motor hand area to a knob on the precentral gyrus. A new landmark. Brain J. Neurol. 120, 141–157.

Zeiler, M.D., Fergus, R., 2013. Visualizing and Understanding Convolutional Networks. ArXiv13112901 Cs.

Zeki, S., Watson, J., Lueck, C., Friston, K.J., Kennard, C., Frackowiak, R., 1991. A direct demonstration of functional specialization in human visual cortex. J. Neurosci. 11, 641–649.

Zhang, L., Dong, W., Zhang, D., Shi, G., 2010. Two-stage image denoising by principal component analysis with local pixel grouping. Pattern Recognit. 43, 1531–1549.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595.

Zhang, Y., Han, K., Worth, R., Liu, Z., 2019. Connecting concepts in the brain: mapping cortical representations of semantic relations. bioRxiv 649939.

Zvonik, E., Cummins, F., 2003. The effect of surrounding phrase lengths on pause duration, in: Eighth European Conference on Speech Communication and Technology.

# NEDERLANDSE SAMENVATTING

In dit proefscrift onderzochten wij de neurale dynamiek tijdens taalverwerking bij het toepassen van datagestuurde modellering in de neurale responsen tijdens het kijken van een film. Wij zagen dat verschillende akoestische, visuele en taalgelateerde eigenschappen voorkomen uit ons bottom-up onderzoek en modellering van de neurale data. Met deze computationele aanpaken konden wij datagestuurd bewijs aandragen voor enkele invloedrijke taalverwerking theorieën. Tegelijkertijd maakten wij nieuwe bevindingen die sommige hedendaagse theorieën uitdagen en ons algemene begrip van de corticale mechanismen die ten grondslag liggen aan taalverwerking uitbreiden.

In **hoofdstuk 2** hebben we low-level akoestische spraakeigenschappen passend gemaakt bij de neurale ECoG-reacties tijdens continue spraakperceptie in een filmexperiment. Het passend maken was succesvol voor veel ECoG-elektroden langs de perisylvische cortex. We vonden bewijs van informatieverspreiding langs de perisylvische cortex tijdens spraakperceptie met gyrus temporalis superior posterior temporele gebieden afgestemd op snelle temporele informatie en gyrus temporalis superior anterior temporele gebieden en gyrus frontalis inferior afgestemd op trage temporele informatie in spraak. Deze waargenomen verschillen in temporele afstemming vielen in verschillende snelheden van hiërarchische taaleenheden: van fonemen tot woorden.

In **hoofdstuk 3** hebben we een volledig datagestuurde benadering ontwikkeld om niet-lineaire auditieve kenmerken uit het onbewerkte geluidssignaal te extraheren op een manier die de bijbehorende corticale reacties het beste verklaart. Hiervoor hebben we een diep kunstmatig neuraal netwerk (ANN) getraind op de onbewerkte soundtrack van een lange speelfilm om de ECoG-gegevens te voorspellen die zijn verzameld tijdens het bekijken van de film. Onze bevindingen toonden dat ons niet-lineair neuraal coderingsmodel bij alle proefpersonen succesvol paste. Belangrijk was dat het het het beste paste in vergelijking met verschillende controle functiesets. Ons model generaliseerde nieuwe neurale- en audiogegevens goed en de bottom-up functies die door ANN werden geëxtraheerd boden datagestuurde ondersteuning voor het idee van informatiepropagatie langs de perisylvische cortex zoals beschreven in hoofdstuk 2.

In **hoofdstuk 4** hebben we gebruik gemaakt van ECoG-opnames met een hoge dichtheid waarmee we de functionele mapping in de sensorimotorische cortex in een hoge mate van ruimtelijk detail konden onderzoeken. We waren in staat om een apart gebied in de voorste dorsale precentrale cortex te identificeren die betrokken was bij het volgen van waargenomen spraak. Dit gebied volgde verschillende perceptueel relevante auditieve eigenschappen van spraak, inclusief - maar niet beperkt tot de spectral envelope van spraak. Belangrijk is dat we hebben aangetoond dat het geïdentificeerde gebied verschilde van de aangrenzende handmotorische en gezichtmotorische regio's.

In **hoofdstuk 5** hebben we ons gericht op de neurale codering van semantische kenmerken verbonden met het waarnemen van een speelfilm. We gebruikten een combinatie van visuele objectherkenning en taalmodel ANN's om semantische informatie te extraheren horende bij verschillende fragmenten van de film. De semantische kenmerken die door ANN's werden geëxtraheerd legden belangrijke semantische onderscheidingen vast, zoals aanwezigheid of afwezigheid van mensen in de film, aanwezigheid of afwezigheid van menselijke gezichten, een statische scène versus momenten met actie en beweging. Deze belangrijkste semantische verschillen zijn in kaart gebracht op de neurale activiteit binnen verschillende verdeelde corticale netwerken.

In het algemeen hebben we in dit werk aangetoond dat de toepassing van een datagestuurde benaderingen voor het modelleren van de neurale reacties op taal en spraak een veelbelovende aanpak is die ons inzicht in de neurale mechanismen van spraak en taal kan bevorderen. We konden aantonen dat de datagestuurde bevindingen nuttig kunnen zijn bij het verfijnen van de huidige neurobiologische theorieën over taalperceptie. Wij geloven dat de gebruikte methoden potentieel hebben voor het bevorderen van zowel theoretische als toegepaste neurowetenschappen door goed presterende computermodellen van de hersenreacties op spraak te leveren.

# SUMMARY IN RUSSIAN

Человеческий мозг уникален не только потому, что является центром нашего мышления и сознания, но и потому что обслуживает уникальную человеческую способность – понимать и использовать речь. Многие виды животных, включая обезьян, дельфинов и птиц, способны на довольно сложные виды общения между собой. Многие животные умеют считать, хоть и в общих чертах (например, отличать один предмет от двух отдельных предметов и от большего числа предметов). Многие животные могут сообщить родственным животным об опасности или о местонахождении еды. Многие используют вокализацию, мимику и телодвижения для общения между собой. Однако только человек как вид способен использовать речь для абстрактного мышления, создавать искусство с помощью речи, формулировать свое понимание мира в науке и описывать историю событий на понятном ему языке.

Сам по себе речевой сигнал – это последовательность звуков, не являющаяся самодостаточной для ее понимания. Например, если вы услышите речь на незнакомом вам языке, вы не сможете ее понять. Правильная обработка услышанного звукового сигнала вашим головным мозгом с учетом вашей памяти и языковой компетенции – ключ к пониманию и использованию языка. Каким образом человеческий мозг, физический орган, состоящий из нервных клеток и кровяных сосудов, преобразует слышимый звуковой сигнал в понятную речь, является одним из наиболее интересных и сложных вопросов в нейробиологии.

В этой работе мы исследовали, как человеческий мозг обрабатывает речь. Для этого мы провели несколько экспериментов, в которых участники слушали речь на родном для них языке. Для этого им был предложен просмотр художественного фильма. Во время просмотра мы регистрировали активность головного мозга у каждого участника. Поскольку наша работа является частью большего проекта, связанного с патологией мозга, нашими участниками были пациенты, проходящие временное наблюдение и лечение от эпилепсии (патология мозга, сопровождающаяся судорожными приступами). Частью этого процесса является оперативное вмешательство, в результате чего на поверхность головного мозга пациента помещается сеть электродов для регистрации мозговой активности. Эта активность наблюдается медицинским персоналом для установления правильного лечения для каждого индивидуального пациента. Регистрация мозговой активности продолжается около недели, после чего в ходе повторной операции сеть электродов удаляется, и пациент проходит лечение от эпилепсии. В течение недели, когда активность мозга постоянно регистрируется через внедренные электроды, мы имеем возможность провести эксперименты с пациентом. В этот момент мы и предлагали пациентам просмотреть художественный фильм.

Таким образом, на первом этапе работы мы собрали данные о мозговой активности каждого пациента, вызванной просмотром фильма (а значит и прослушиванием его звуковой дорожки, содержащей диалоги на родном пациенту языке). После этого мы анализировали собранные данные. Разные главы данной работы описывают различные аспекты нашего анализа. Например, во второй главе мы описываем физические свойства речевого сигнала (такие как амплитуда и фаза колебаний звуковой волны). Мы также говорим о том, что несколько областей головного мозга, особенно в височной и лобовой коре, коррелируют со звуковыми свойствами речи. Мы видим, что это происходит в большей степени для речи, на которую пациенты активно обращают внимание. Звуки фильма, которые привлекают меньше внимания участников эксперимента, ведут к меньшей активности мозга и меньшей корреляции с физическими свойствами звука.

В другой главе данной работы (глава 5) мы описываем, как с помощью современных методов анализа данных мы может исследовать, какие области мозга содержат смысловую информацию о просмотренном (и прослушанном) фильме. Это может быть информация о главных героях фильма и их действиях, моменты диалогов между героями, а также полное отсутствие людей на экране в некоторых сценах. Фактически, нам удается прочитать зарегистрированную мозговую активность наших пациентов и правильно угадать, к какому моменту фильма она относится. Эти результаты позволяют нам лучше понять, как смысл услышанной речи отражается в активности головного мозга.

В целом, данная работа показывает, что разные аспекты услышанной речи, такие как акустические свойства звука, ее громкость, тон, ритм, а также смысловое наполнение, отражены в зарегистрированном сигнале головного мозга. Мы видим, что определенные зоны мозга (височная и лобная части) больше связаны с обработкой физической стороны речи, в то время как весь мозг вовлечен в обработку смысла речи. Эти результаты позволяют нам лучше понять, что происходит в мозге человека, когда он воспринимает речь других. Мы надеемся, что наша работа не только будет полезна для научного сообщества, но также привлечет более широкую аудиторию и заинтересует ее уникальной способностью головного мозга человека воспринимать речь.

## ACKNOWLEDGEMENTS

I would like to thank many people who have helped me in creating this manuscript. Many thanks go to my colleagues at the University Medical Center Utrecht. In particular, I thank my co-promoter Zac Freudenburg, who guided me throughout my entire PhD by brainstorming, discussing, analyzing and interpreting our results. I am also thankful for his sharing of my passion for music and soccer, the jamming of the High Gamma band and all the little chats we have had over this time. Next, I thank Mariska Vansteensel, who organized most of the patient-related work in the lab; I would not have come far in this thesis without her help with obtaining the data; Erik Aarnoutse, who provided us with technical support at every level, including data acquisition and the knowledge database; Mariana Branco, who developed better, automated pipelines for data preprocessing and data structuring; Efraïm Salari, with whom we discussed our ECoG/speech-related findings and tried to make sense of our data; Sacha Leinders, who helped with many technical ECoG issues; Mathijs Raemaekers, who helped with all fMRI questions; Mark Bruurmijn, who developed EcoG analyses tools; Meron Vermaas, with whom I often chatted about our PhD hardships; Anouck Schippers, who shares my passion for language and ECoG; my officemate Giovanni Piantoni, who also prefers Python over Matlab and my former officemate Fransisco Fernandes, who installed a 'cookie jar' in our office. Many other people contributed to this work through discussions, by helping with the workflow and practical issues, and by simply making it a great place to work at: Wouter Schellekens, Natalia Petridou, Maxime Verwoert, Miek de Dreu, Anoek Smits, Jinne Geelen, Anna Gaglianese, Elmar Pels, Philippe Cornelisse, Benny Van Der Vijgh. Special thanks go to Dora Hermes and Kai Miller for their ECoG expertise and invaluable advice on interpretation and visualization of the data. I would like to thank Vincent Kersten for maintaining our work servers and keeping our data safe. Special thanks go to Annemiek Elbertse, who always helped me out with a lot of confusing and tedious paperwork and who made my transition from a PhD candidate to a guest postdoctoral researcher smoother than I imagined. I would also like to thank my students Zoe Cox-Putker, Daniëlle Bruel, Clarissa Baratin and Anastasia Nosanova, whom I have genuinely enjoyed supervising and who all did great on their final reports.

Very special thanks go to my supervisor Nick Ramsey. At the end of 2014 I applied for several PhD positions, but hopefully some people (Luca?) can still confirm that the job offer at prof. Ramsey's BCI group was my very favorite application. In fact, I badgered Annemiek and Nick for months, before the position even became available indicating my boundless enthusiasm for it. Then I almost blew my chances of getting it by arriving hopelessly late for my job interview. Lucky for me though, Nick seemed to be forgiving of my punctuality-related shortcomings and just a couple of weeks before Christmas 2014 he offered me to join his team. Throughout my PhD he has been a fantastic supervisor, offering me freedom in finding and making my own projects yet always keeping an eye on the 'big questions' and the 'big picture'. He encouraged me to pursue challenging topics, long-range collaborations and ambitious journal submissions. I enjoyed every work-related and off-topic conversation with him. We have talked about the speech representation in the brain, the future of BCI, the future of the scientific community, peer-review process, data sharing, grants and funding, the future of humanity, progress in wartime, vegetarianism, sailing, retirement and the best pizza in the world. I am so lucky that I have had a supervisor who could truly inspire me to do my job. From the bottom of my heart I want to thank him for everything he did for me and said to me in these past five years. I could not have asked for a better supervisor than he has been to me. I hope that we will continue our work as collaborators and will do more great research together.

Next, I want to thank my colleagues at the Donders Center for Brain, Cognition and Behaviour, and in particular, my long-time collaborator and promotor of this work – Marcel van Gerven. I came to the Netherlands for a master's internship in 2013 because I saw online the wonderful work of his lab on decoding images from the brain activity. In many ways this transition shaped my future research plans. I am thankful for the experience we have had so far and I am very excited about the experience that is coming with my postdoctoral position in his lab. I want to thank another collaborator from the Donders center – Umut Güçlü, who helped with planning and interpreting the computational modeling part of this thesis. Next is Luca Ambrogioni, with whom I had many insightful brainstorming sessions about the modeling and the interpretation of the results in the present work. Luca is not only a valuable collaborator, but is also a dear friend. And because he is, I am not holding his interesting take on the 'acknowledgments' in his thesis against him. Also, this is as good a place as any to admit that I got a 7.0 on that exam he tutored me for. Shameful perhaps, but what can you do. I admit that there is room for improvement and I am happy to work on it.

I would also like to thank many of my friends for being there for me during the ups and downs of this entire time of working towards my PhD. I thank all the theater groups I have played with in this time, along with everyone I met at the horse riding school (including the horses), everyone at the pole dancing schools, various gyms, miscellaneous social events as well as trips, leisure or work-related. In that particular regard, I am thankful to everyone I met during summer training programs and scientific conferences. I have been lucky to meet amazing people, who not only shared their interesting research with me, but engaged me in various discussions leading to several collaboration projects. Many of them directly influenced the work presented here.

I also want to thank my family for their love and support. None of this work would have been possible without my mom's gentle encouragement and her strong belief that acting is not a good career choice. I am very happy with where I am in my life and I heartily thank my mom for it. I thank my stepdad and my grandmother too. All of them shaped me as a person. I think I got my independence and determination from both my grandmother and my mom. My mom has always amazed me with her inner strength, passion for what she does and generosity towards the people she loves. My grandmother taught me that most rules are basically just suggestions and everyone makes their own path. My stepdad taught me how to drill a wall and install a ceiling. I sincerely thank him for helping me get out of my head sometimes. I also thank him for all the board games we have played together, his jokes and light spirit no matter the circumstances. Finally, I thank my dad as well. It pains me that he is not here to celebrate this moment of my life. I know that he would have been proud as he always was and I thank him for everything he ever did for me.

/Мне хотелось бы отдельно поблагодарить маму, Диму и бабушку Соню. Эта работа не увидела бы свет без их поддержки. Моя мама всегда считала, что актерская карьера – это не очень удачный выбор профессии. Я очень счастлива в том, где нахожусь сейчас и чем занимаюсь, поэтому думаю, что мама была права. Несмотря на то, что мне не часто удается видеться с семьей, каждый из них оказал огромное влияние на меня. Я думаю, что моя профессиональная и личная независимость перешла мне от обеих мамы и бабушки. Мне кажется, от бабушки мне еще досталась способность не переживать о мнении окружающих, быть себе на уме и уметь выбраться совершенно из любой ситуации с высоко поднятой головой. От мамы, хочется надеяться, мне пришло почти все остальное: сила характера, самоотдача и необъятная щедрость к самым близким людям. Почти все, кроме безвозмездной отдачи своей работе. Мне бы хотелось, чтобы она берегла себя больше. Кроме мамы и бабушки я бы еще хотела поблагодарить маминого мужа Диму. За то, что он научил меня сверлись стены и класть потолок. В моей научной карьере очень важно уметь отвлечь голову и занять руки для разнообразия. Кроме этого, большое спасибо ему за настольные игры, светлый юмор и заразительный оптимизм. Наконец, я бы еще хотела поблагодарить своего папу. Мне сложно передать, как мне жаль, что он не разделит достижения публикации этой книги со мной. Но я знаю, что он был бы очень горд и счастлив за меня, и я благодарна за все, что он когда-либо для меня сделал. Большое спасибо ему и всей моей семье!/

Last but not least, I want to thank Max, not only for making the beautiful cover of this thesis (and the Dutch translation of the summary), but for everything else he does. It was his presence in my life that gave me the resources to push through and write three quarters of this work within a year. It was a tough but very happy year indeed. Thank you.

# CURRICULUM VITAE

## Experience

Postdoctoral researcher, Department of Artificial Intelligence, Donders Institute for Brain, Cognition and Behaviour, Radboud University (February 2020 – present)

Ph.D. candidate, University Medical Center Utrecht, Brain Center Rudolf Magnus, Utrecht, the Netherlands (March 2015 – January 2020)

## Education

Ph.D. Cognitive Neuroscience from University Medical Center Utrecht, Brain Center Rudolf Magnus, Utrecht, the Netherlands (March 2015 – January 2020)

M.Sc. Cognitive Neuroscience from University Of Trento, Italy (September 2012 – September 2014)

Specialist's degree in Linguistics from Lomonosov Moscow State University, Moscow, Russia (September 2007 – June 2012)

## Awards and fellowships

FENS travel grant for the SfN Neuroscience meeting (2019)
Student award for the BCI meeting (2018)
Premio di Merito, University of Trento (2014)
Erasmus Placement 'Place your Talent' Scholarship, University of Trento and University of Venice (2014)
Erasmus Consortium Placement Scholarship, University of Trento (2013)
Erasmus Study Scholarship, University of Trento (2013)
Presidential scholarship, Lomonosov Moscow State University (2007)

## Publications

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A., & Ramsey, N. F. Brain-optimized extraction of complex sound features that drive continuous auditory perception. *In review.*

Berezutskaya, J., Baratin C., Freudenburg, Z. V. & Ramsey, N. F. High-density intracranial recordings reveal a distinct site in anterior dorsal precentral cortex that tracks perceived speech. *In review.*

Berezutskaya, J., Freudenburg, Z. V., Ambrogioni L., van Gerven, M. A., & Ramsey, N. F. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. *In review.*

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M.A.J., Ramsey, N. F. (2017). Neural tuning to low-level features of speech throughout the perisylvian cortex // Journal of Neuroscience, 37(33), p. 7906.

Berezutskaya, J., Freudenburg, Z. V., Ramsey, N. F., Güçlü, U., van Gerven, M.A.J. (2017) Modeling brain responses to perceived speech with LSTM networks. // Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning, p. 149.

Ambrogioni, L., Berezutskaya, J., Güçlü, U., van den Borne, E. W., Güçlütürk, Y., van Gerven, M.A.J., & Maris, E. G. G. (2017). Bayesian model ensembling using meta-trained recurrent neural networks // Workshop on Meta-Learning at Neural Information Processing Systems, p. 1.

Pechenkova E., Vlasova R., Berezutskaya J., Sinitsyn V. (2012) One-back task functional localizer for visual word form area reveals inverse pattern of activation in readers of Russian // Journal of Vision. Vol. 12, n. 9, p. 531.

## Teaching

Berezutskaya, J. Fundamentals of machine learning applied to BCI research at Fundamental didactic session of BCI Society meeting, Asilomar, USA, 2018

Berezutskaya, J. Deep learning for neuroscience at mini-course on machine learning, Brain Center Rudolf Magnus, Utrecht, the Netherlands, 2017

Berezutskaya, J., Branco M.P. MATLAB course in Master's Neuroscience and Cognition program at University of Utrecht, the Netherlands (2017 – 2019)

## Conference Abstracts

Berezutskaya J., Freudenburg Z.V., Aarnoutse E.J., Vansteensel M.J., Leinders S., Pels E., Ramsey N.F. Using a convolutional neural network for improved click detection in an implanted BCI setup at BCI Society meeting, Asilomar, USA, 2018 (talk + poster)

Baratin C., Berezutskaya J., Ramsey N.F. Motor cortex activity in response to perceived speech at Mind the Brain Symposium, Utrecht, the Netherlands, 2018 (poster)

Berezutskaya J., Freudenburg Z.V., Ramsey N. Decoding individual phonemes from ECoG neural responses using convolutional neural networks at Society for Neuroscience, Washington DC, USA, 2017 (poster)

Freudenburg Z.V., Berezutskaya J., Vansteensel M.J., Aarnoutse E.J., Ramsey N.F. Asynchronous detection and classification of spoken phonemes using mouth motor neural correlates in high-density ECoG at Society for Neuroscience, Washington DC, USA, 2017

Berezutskaya J., Freudenburg Z.V., Ramsey N.F., Güçlü, U., van Gerven M.A.J. Modeling continuous ECoG responses to naturalistic speech using recurrent neural networks at Human Brain Mapping, Vancouver, Canada, 2017 (talk)

Berezutskaya J., Freudenburg Z.V., Güçlü, U., van Gerven M.A.J., Ramsey N.F. Neural tuning to low-level features of speech in the brain at Society for Neuroscience, San Diego DC, USA, 2016 (talk)

Cox-Putker Z., Berezutskaya J., Ramsey N.F. Understanding encoding of complex visual features in a continuous feature film at Mind the Brain Symposium, Utrecht, the Netherlands, 2016 (poster)

Bruel D., Berezutskaya J., Ramsey N.F. Encoding of perceived speech intensity in the brain using fast fMRI at Mind the Brain Symposium, Utrecht, the Netherlands, 2016 (talk)

Berezutskaya J., Freudenburg Z.V., Ramsey N.F., Güçlü, U., van Gerven M.A.J. Low-level encoding of continuous speech in high-gamma neural responses at Donders Discussions, Nijmegen, the Netherlands, 2015 (poster)

Berezutskaya J., Simanova I., van Gerven M.A.J. Representation of mammals in the human brain at Society for Neurobiology of Language, Amsterdam, the Netherlands, 2014 (poster)

Litvinova L.D., Pechenkova E.V., Vlasova R.M., Berezutskaya J., Sitinsin V.E. Localization of speech perception using three experimental paradigms (fMRI evidence from Russian) at International Symposium on Functional Neuroimaging: Basic Research and clinical Applications, Moscow, Russia, 2012 (poster)

Berezutskaya J., Pechenkova E.V. Localization of semantic and syntactic processing using functional magnetic resonance imaging (evidence from Russian) at Neuro- and Psycholinguistic Approaches to Language Processing, Braga, Portugal, 2011 (poster)