Analysis of cortical activity for the development of brain-computer interfaces for speech

Philémon Roussel

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement

Arrêté ministériel : 25 mai 2016

Présentée par

Philémon ROUSSEL

Thèse dirigée par **Blaise YVERT**, Directeur de recherche, INSERM

préparée au sein du **Laboratoire Grenoble Institut des Neurosciences**
dans **l'École Doctorale Ingénierie pour la Santé la Cognition et l'Environnement**

Analyse d'activité corticale pour le développement d'interfaces cerveau-machine pour la parole

Analysis of cortical activity for the development of brain-computer interfaces for speech

Thèse soutenue publiquement le **11 juin 2021**,
devant le jury composé de :

Monsieur BLAISE YVERT
DIRECTEUR DE RECHERCHE, INSERM DELEGATION AUVERGNE-RHONE-ALPES, Directeur de thèse
Madame TETIANA AKSENOVA
CADRE SCIENTIFIQUE DES EPIC, CEA CENTRE DE GRENOBLE, Examinatrice
Monsieur STEPHAN CHABARDES
PROFESSEUR DES UNIV - PRATICIEN HOSP., UNIVERSITE GRENOBLE ALPES, Président
Madame ANNE-LISE GIRAUD
PROFESSEUR, Université de Genève, Examinatrice
Monsieur FRANK GUENTHER
PROFESSEUR, Boston University, Examinateur
Monsieur BENJAMIN MORILLON
CHARGE DE RECHERCHE HDR, INSERM PACA - CORSE, Rapporteur
Monsieur WILSON TRUCCOLO
PROFESSEUR ASSOCIE, Brown University, Rapporteur

Résumé

Les interfaces cerveau-ordinateur pour la parole pourraient permettre de restaurer la faculté de parler chez des personnes souffrant de paralysie sévère. Un tel système décoderait des caractéristiques de ce que l'utilisateur souhaite prononcer à partir de son activité cérébrale et synthétiserait les sons associés en temps réel. De récentes études montrent des résultats prometteurs en ce sens, en réalisant des décodages à partir d'enregistrements cérébraux effectués chez des participants sains parlant à voix haute. Cependant, plusieurs défis doivent être relevés pour permettre le décodage de la parole chez des personnes dans l'incapacité de parler. Cette thèse vise à contribuer au développement des interfaces cerveau-ordinateur ainsi qu'à la compréhension de l'activité corticale sous-tendant la parole.

Premièrement, nous avons mis en évidence un phénomène de contamination acoustique des signaux électrophysiologiques pendant la parole. Nos analyses ont révélé que ce phénomène était présent dans plusieurs enregistrements, provenant de différents laboratoires dans le monde, et qu'il pouvait avoir un impact important sur les études de décodage de la parole en biaisant les performances obtenues. Nous avons étudié l'origine de ce phénomène et proposé une méthodologie pour le quantifier.

Il est attendu que l'activité corticale de sujets paralysés essayant de parler soit partiellement différente de celle de sujets sains s'exprimant à voix haute, en particulier à cause de l'absence de retours auditif et somatosensoriel. Comprendre les différences entre ces deux conditions est important pour savoir si les modèles actuellement développés sur la parole à voix haute seraient facilement applicable à des personnes paralysées. Dans cette optique, étudier la parole imaginée chez des participants sains permet d'étudier l'activité sous-tendant la parole en l'absence de retours sensoriels, comme c'est le cas chez les personnes paralysées.

A partir d'un enregistrement intracortical, nous avons montré que la dynamique de l'activité d'une population de neurones dans l'aire de Broca était différente pendant la parole à voix haute et imaginée. De plus nous avons trouvé que l'arrangement temporel des potentiels d'action enregistrés pendant la prononciation de phrases permettait de discriminer ces phrases. La similarité des trains de potentiels d'action s'est avérée être corrélée à la similarité phonologique des phrases. Enfin, à partir du même enregistrement, nous avons montré que la composante basse fréquence de potentiels de champ locaux était modulée par la parole à voix haute et dans une moindre mesure par la parole imaginée.

Un autre obstacle concernant le décodage de la parole interne chez des participants en incapacité de parler est la difficulté de calibrer le décodeur. En effet, en l'absence de signaux comportementaux il est n'est alors pas possible de connaitre précisément le timing auquel le participant imagine parler. Nous avons donc testé, chez un participant non-paralysé, un protocole permettant de construire un décodeur de parole imaginée à partir d'activité électrophysiologique. Nos analyses a posteriori ont montré qu'il était possible de décoder des caractéristiques acoustiques de phonèmes imaginés avec des performances statistiquement supérieures à la chance. Nous avons également réalisé chez le même participant un test préliminaire de décodage en temps réel et en boucle fermée.

Abstract

Brain-computer interfaces for speech have the potential to restore the ability to speak in people with severe paralysis. Such a system would decode features of what the user wants to say from their brain activity and synthesize the associated sounds in real time. Recent studies show promising results in this direction by decoding speech from brain recordings of healthy participants speaking aloud. However, several challenges must be overcome to enable speech decoding in people unable to speak. This thesis aims to contribute to the development of brain-computer interfaces and to the understanding of cortical activity underlying speech.

First, we highlighted a phenomenon of acoustic contamination of electrophysiological signals during speech. Our analyses revealed that this phenomenon was present in several recordings from different laboratories around the world and that it could have an important impact on speech decoding studies by biasing the obtained performances. We studied the origin of this phenomenon and proposed a methodology to quantify it.

It is expected that the cortical activity of paralyzed subjects attempting to speak is partially different from that of healthy subjects speaking aloud, in particular because of the absence of auditory and somatosensory feedback. Understanding the differences between these two conditions is important to know if the models currently developed on overt speech would be easily applicable to paralyzed people. In this perspective, studying covert speech in healthy participants allows us to study the activity underlying speech production in the absence of sensory feedback, as it is the case for people with paralysis.

Using an intracortical recording, we showed that the dynamics of the activity of a population of neurons in Broca's area were different during overt and covert speech. Moreover, we found that the temporal arrangement of action potentials recorded during the pronunciation of sentences allowed to discriminate these sentences. The similarity of spike trains appeared to be correlated with the phonological similarity of the sentences. Finally, using the same recording, we showed that the low frequency component of local field potentials was modulated by overt speech and to a lesser extent by covert speech.

Another obstacle concerning the decoding of attempted speech in participants unable to speak is the difficulty to calibrate the decoder. Indeed, in the absence of behavioral output, it is not possible to know precisely the timing at which the participant attempts to speak. We therefore tested, in a non-paralyzed participant, a protocol allowing to build a covert speech decoder from electrophysiological activity. Our offline analyses showed that it was possible to decode acoustic features of imagined phonemes with performances statistically superior to chance. We also performed a preliminary test of real-time, closed-loop decoding in the same participant.

Acknowledgments

First of all I would like to thank Blaise Yvert, who supervised me throughout this thesis, for his availability, his benevolence and his precious teachings. He made it possible for me to elaborate and complete this work by offering his ideas and his experience and also by sharing his energy and unfailing enthusiasm for research.

I thank Stéphan Chabardès for his role as president of the jury and for his indispensable involvement in the acquisition of the data on which my thesis is based. I also thank all the members of the jury, in particular the reviewers, for taking the time to evaluate my thesis and for their insightful comments.

I am sincerely grateful to ALL the members of the team, who helped me all along through their advice, feedbacks, encouragements and the friendships we built. I have of course a special thought for Marie Palma and Gaël Le Godais without whom life in the lab would not have been the same. I also address special thanks to Florent Bocquelet for everything he taught me before and during the thesis.

Finally, I cannot fail to thank my family and friends who have supported me in so many ways throughout the completion of this work.

Contents

Acronyms **12**

List of Figures **15**

List of Tables **17**

Introduction **19**
Context and motivation 19
Objectives of the thesis 20
Organization of the manuscript 21

1 Brain-computer interfaces **23**
1.1 Principles 23
1.2 Brain activity recording 23
1.3 Feedback and adaptation 26
1.4 Brain-computer interfaces for communication 28

2 Quantifying and synthesizing speech **29**
2.1 Physical characteristics 29
2.1.1 Phonemes 29
2.1.2 Speech production mechanisms 29
2.1.2.1 Vocal cords 30
2.1.2.2 Articulators 30
2.1.2.3 Measuring articulatory movements 30
2.1.2.4 Inferring articulation from sound 31
2.1.3 Acoustics 31
2.1.3.1 Spectro-temporal structure 31
2.1.3.2 Acoustic representations 32
2.2 Semantic representations of speech 33
2.3 Speech synthesis 34
2.3.1 Acoustic synthesis 34
2.3.2 Articulatory synthesis 34
2.3.3 Text-to-speech synthesis 35

3 Neural basis of speech production **39**
3.1 Cortical organization of speech production . . . 39
3.1.1 The dual-stream model of speech processing . . . 39
3.1.2 Roles of frontal areas in speech production . . . 41
3.1.2.1 Inferior frontal gyrus . . . 41
3.1.2.2 Ventral sensorimotor cortex . . . 42
3.2 Models of speech motor control . . . 43
3.3 Covert speech . . . 45

4 Decoding speech production from electrophysiological activity **47**
4.1 Speech neural features . . . 47
4.1.1 Firing rates . . . 47
4.1.2 Local field potentials . . . 48
4.2 Dimensionality reduction and features selection . . . 48
4.3 Speech decoding models . . . 50
4.3.1 Discrete decoding . . . 50
4.3.2 Continuous decoding . . . 51
4.4 Review of speech decoding performances . . . 52
4.4.1 Discrete decoding . . . 52
4.4.2 Continuous decoding . . . 54

5 Acoustic contamination of electrophysiological recordings **55**
5.1 Introduction . . . 55
5.2 Neural datasets . . . 56
5.2.1 Brainspeak datasets . . . 56
5.2.1.1 Participants . . . 56
5.2.1.2 Electrophysiological recordings . . . 56
5.2.1.3 Audio recordings . . . 58
5.2.1.4 Tasks and stimuli . . . 58
5.2.1.5 Closed-loop recordings . . . 58
5.2.2 Additional human datasets . . . 59
5.2.2.1 Participants . . . 59
5.2.2.2 Electrophysiological recordings . . . 59
5.2.2.3 Audio recordings . . . 60
5.2.2.4 Tasks and stimuli . . . 60
5.2.3 Rat recording . . . 61
5.3 Methods . . . 61
5.3.1 Pre-processing . . . 61
5.3.1.1 Audio preprocessing . . . 61
5.3.1.2 Segments selection . . . 62
5.3.1.3 Channel selection . . . 62
5.3.1.4 Robust estimation of standard deviation . . . 62
5.3.1.5 Artifacts exclusion . . . 62
5.3.1.6 Common average reference . . . 63
5.3.2 Spectrogram computation . . . 63

5.3.3 Audio-neural correlations . . . 63
5.3.4 Objective assessment of contamination . . . 64
5.3.5 Neural decoding . . . 64
5.4 Results . . . 65
5.4.1 Observation of acoustic contamination in neural recordings . . . 65
5.4.1.1 Correlation between ECoG and sound signals during speech production . . . 65
5.4.1.2 Correlation between intracortical and sound signals during speech production . . . 65
5.4.1.3 Correlation between electrode and sound signals during sound perception . . . 67
5.4.1.4 Audio-neural cross-correlations . . . 68
5.4.1.5 Objective assessment of acoustic contamination . . . 68
5.4.2 Potential influence of contamination on speech decoding . . . 71
5.4.3 Possible sources of acoustic contamination . . . 74
5.4.3.1 Sound contamination and electrode quality . . . 74
5.4.3.2 Electrode versus connector mapping of contamination . . . 76
5.4.3.3 In vitro evidence of acoustic contamination . . . 77
5.4.3.4 Localization of acoustic contamination along the recording chain . . . 80
5.5 Discussion . . . 80

6 Dynamics of neural activity in Broca's area during overt and covert speech 85
6.1 Introduction . . . 85
6.2 Material and methods . . . 86
6.2.1 Micro-electrode recording . . . 86
6.2.2 Protocol . . . 86
6.2.3 Trial phases labeling . . . 87
6.2.4 Neural signal processing . . . 88
6.2.4.1 Common average reference . . . 88
6.2.4.2 Spike-sorting . . . 88
6.2.4.3 Firing rates . . . 88
6.2.4.4 Slow LFPs . . . 90
6.2.5 Condition classification . . . 90
6.2.5.1 Labeling . . . 90
6.2.5.2 Features . . . 90
6.2.5.3 Linear discriminant analysis . . . 90
6.2.5.4 Mathews correlation coefficient . . . 91
6.2.5.5 Phoneme segmentation . . . 91
6.2.6 Spike train analysis . . . 91
6.2.6.1 Spike train distance . . . 91
6.2.6.2 Spike train classification . . . 93
6.2.7 Sentences similarity . . . 93
6.2.7.1 Phonemic distance . . . 93

6.2.7.2 Semantic distance . . . 93
6.2.7.3 Possible confounders . . . 94
6.2.8 Spatial mapping . . . 95
6.2.9 Relationship between LFP phase and firing rates . . . 96
6.3 Results . . . 96
6.3.1 Speech-related modulation of firing rates . . . 96
6.3.1.1 Single unit and population modulation . . . 96
6.3.1.2 Classification . . . 96
6.3.1.3 Rate coding of phonemes . . . 100
6.3.2 Spike train similarities during speech . . . 100
6.3.2.1 Classification of sentences based on spike train distances 101
6.3.2.2 Relationship between spike train distances and sentence similarity metrics . . . 103
6.3.2.3 Spike train similarities during covert speech . . . 105
6.3.3 Speech-related modulation of slow LFPs . . . 106
6.3.3.1 Slow LFPs during overt speech . . . 106
6.3.3.2 Slow LFPs during covert speech . . . 108
6.3.3.3 Relationship between firing rates and slow LFP phase . 109
6.4 Discussion . . . 110

7 Preliminary decoding of covert speech based on electrocorticography 113
7.1 Introduction . . . 113
7.2 Material and methods . . . 114
7.2.1 Subject . . . 114
7.2.2 Set of phonemes . . . 114
7.2.3 BY2014 articulatory-acoustic corpus . . . 115
7.2.4 Acoustic and articulatory representations of phonemes . . . 115
7.2.5 Offline open-loop decoding . . . 116
7.2.5.1 Protocol . . . 116
7.2.5.2 Task labeling . . . 117
7.2.5.3 Neural data processing . . . 117
7.2.5.4 Estimation of acoustic and articulatory representations of the imagined phonemes . . . 117
7.2.5.5 Regression . . . 118
7.2.5.6 Features selection . . . 118
7.2.5.7 Cross-validation . . . 118
7.2.5.8 Classification of the decoded acoustic representation . . 118
7.2.5.9 Chance level estimation . . . 119
7.2.6 Online closed-loop decoding . . . 119
7.2.6.1 Protocol . . . 119
7.2.6.2 Software . . . 119
7.2.6.3 Features extraction . . . 120
7.2.6.4 Closed-loop calibration . . . 122
7.2.6.5 Decoding and synthesis . . . 122
7.3 Results . . . 124

7.3.1 Band power modulation during covert speech 124
7.3.2 Offline open-loop decoding 124
7.3.2.1 Influence of features selection 126
7.3.2.2 Decoding of acoustic and articulatory representations . 126
7.3.2.3 Contribution of the different frequency band powers of the LFP 128
7.3.2.4 Contribution of the electrodes 128
7.3.2.5 Decoding using low-passed LFP amplitude 128
7.3.2.6 Phoneme discrimination based on the decoded acoustic representations 128
7.3.3 Online closed-loop decoding 129
7.3.3.1 Technical performances 130
7.3.3.2 Decoding performances 130
7.4 Discussion 130

Conclusions and perspectives **133**

Annexes **137**

Bibliography **139**

Acronyms

ANN Artificial neural network.

BCI Brain-computer interface.

CAR Common average reference.

DNN Deep neural network.

DoF Degree of freedom.

ECoG Electrocorticography.

EMA Electromagnetic articulography.

FR Firing rates.

IFG Inferior frontal gyrus.

LDA Linear discriminant analysis.

LFP Local field potentials.

LOO Leave-one-out.

MAD Median absolute deviation.

MEA Micro-electrode array.

MEL Mel-cepstral coefficients of the MLSA filter.

MFCC Mel-frequency cepstral coefficients.

MLSA Mel Log Spectrum Approximation.

PCA Principal component analysis.

STG Superior temporal gyrus.

vSMC Ventral sensorimotor cortex.

List of Figures

1 Block diagram of a BCI application 24
2 Electrophysiological measurement methods of brain activity 26
3 Categories of BCIs according to the adaptation paradigm 27

4 Illustration of the vocal tract 36
5 Electromagnetic articulography 37
6 Speech formants 37

7 Schematic diagram of the dual-stream model of speech 40
8 Location of the inferior central gyrus and its subregions 42
9 Mapping of encoded articulatory gestures on the vSMC 43
10 Schematic representation of the cortical components of the DIVA model of speech production and acquisition 44

11 Electrode placement for the different human participants 57
12 Correlation between voice and ECoG signals during speech production in participant P2 66
13 Correlations between voice and intracortical signals during speech production in participant P3 67
14 Spectrogram correlations between sound and neural data for participant P5 69
15 Correlations between sound and µ-ECoG spectrograms during pure tones perception in an anesthetized rat 70
16 Audio-neural cross-correlations 71
17 Objective assessment of acoustic contamination of 6 speech production datasets 72
18 Objective assessment of acoustic contamination of 8 sound perception datasets 73
19 Linear decoding of speech features using either contaminated or not contaminated neural recordings 75
20 Correlations between sound and ECoG recordings in participant HG during speech perception 76
21 Spatial organization of acoustic contamination according to the grid and connector layouts for participant P5 during sound perception 78
22 Correlations between sound and ECoG recordings in different in vitro experimental configurations 79

23 Determination of the location of sound contamination along the recording chain . . . 81
24 Localization of the Utah array for participant P3 . . . 86
25 Course of a trial for the speech production experiment performed by participant P3 . . . 87
26 Illustration of the spike-sorting output . . . 89
27 Illustration of the VP distance . . . 92
28 Visualization in 2D of the 100-dimensional vector space of word representations for the experiment of participant P3 . . . 95
29 Condition-modulated units . . . 97
30 3D trajectory of modulated firing rates in PCA space . . . 98
31 Speech condition classification results . . . 99
32 Examples of spike trains and distribution of multi-unit spike train distance for repeated sentences . . . 101
33 Matrix of average multi-unit spike train distances for the different pairs of sentences . . . 102
34 Classification accuracy for different number of units. . . . 103
35 Relationships between the multi-unit spike train distance and the sentence similarity metrics . . . 104
36 Example of slow LFP variations . . . 107
37 Mapping of the average slow LFP on the Utah array during overt and covert repetitions . . . 108
38 Relationship between firing rates and slow LFP phases . . . 109
39 Localization of the ECoG electrodes and Utah array for participant P5 . . . 115
40 Course of a trial for the covert phoneme production experiment performed by participant P5 . . . 116
41 Photograph of participant P5 during the closed-loop experiment . . . 120
42 Screenshot of the GUI of the PulsIO sofwtare used for closed-loop decoding 121
43 Schematic diagram of the online decoder . . . 123
44 Band power modulation during covert speech . . . 125
45 Results of the offline decoding of covert phonemes . . . 127
46 Examples of decoded acoustic coefficients . . . 129

List of Tables

1 Comparison of the brain activity real-time measurement methods. 25

2 Summary of speech decoding studies based on electrophysiological data . 53

3 Assessment of the presence or absence of contamination for 20 datasets . 74

Introduction

Context and motivation

Speech disorders, assuming various forms and causes, affect the ability to communicate of millions of people worldwide. The loss of the ability to speak can result from brain strokes, brainstem trauma as well as from neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS) or Parkinson's disease. Depending on their physiological causes and associated symptoms, speech disorders fall under different categories. In particular, aphasia refers to language disorders (with potential impact on speech) resulting from damage in specific brain regions and dysathria to speech disorders caused by neurological injury of the motor component of the speech production. Associations estimate that aphasia affects 2 000 000 people in the United-States and 300 000 in France[1]. As for dysarthria, studies show that it affects approximately 10 to 60% of patients with stroke or brain injury, 70-100% of patients suffering from Parkinson's disease and almost all patients in later stages of ALS[2]. In the case of ALS, the potential loss of speech is often considered as one of the worst aspects of the disease (Hecht et al., 2002) and most people require augmentative and alternative communication methods within the first years after symptoms onset (Makkonen et al., 2018).

Locked-in syndrome describes the condition of patients with an intact consciousness who suffer from a almost complete loss of voluntary motor control. The extent of the residual motor control varies but is often limited to eye movements and blinking. There are currently more than 500 people in France living with locked-in syndrome, in most cases as a result of a stroke or a trauma[3]. The ability of these people to communicate most often relies on eye movements. During the last decades, alternative means of communication have been developed to help people suffering from locked-in syndrome and other paralysis involving severe dysarthria. The technologies aiming to restore the ability to communicate for patients with speech impairment can rely on residual muscular activity or directly on brain activity. Systems such as spellers based on eye tracking or neck movements provide a reliable way to communicate but remain slow and demanding compared to natural speech and are not usable to all individuals, depending on their condition. Brain-computer interfaces (BCI) for communication remain for now experimental but have the potential

[1]Estimates provided by the National Aphasia Association and the *Fédération Nationale des Aphasiques de France*.

[2]Estimates obtained by the American Speech-Language-Hearing Association from a compilation of studies.

[3]Information provided by the *Association du Locked-In Syndrome*

to allow the most severely impaired individuals to express themselves faster and with less effort. Recent developments in the field of BCI let us envision the possibility of driving a speech synthesizer using features extracted from brain signals in real-time.

Pursuing the goal of developing a BCI for speech is also an opportunity to study the neural mechanisms underlying speech production. In the same way that research on motor BCIs allows to explore the neuronal representations of movement and the organization of motor control in the brain, the attempt to decode speech-related activity provides insights into the fundamental brain dynamics underlying speech production. The general goal of this thesis is to contribute to the development of speech BCIs and to the understanding of brain activity during speech production.

Objectives of the thesis

Most preliminary works on speech BCIs focus on decoding neural activity during overt speech production and recent studies show very promising results in this direction. While achieving overt speech decoding would undeniably constitute a major step, several challenges remain to be solved in order to enable people unable to speak to use a speech BCI. Based primarily on electrophysiological recordings of human participants at the Grenoble-Alpes University Hospital (*Centre Hospitalier Universitaire Grenoble-Alpes*), several of these challenges were adressed.

First of all, the observation of electrophysiological signals revealed that sounds could, in some recordings, disrupt the biopotential measurements. This problem has important implications for studies decoding brain activity during speech production and perception. **Our first objective was to describe this phenomenon of acoustic contamination in neural data recordings, to identify its origin and to propose a method to objectively assess its presence.**

The cortical activity of paralyzed people during attempted speech is expected to be partially different from the one of healthy participants during overt speech, in particular because of the absence of natural auditory and somatosensory feedbacks resulting from the inability to articulate. Understanding the difference between these two conditions is important to know if the models currently developed on overt speech would transfer well to paralyzed people. In this perspective, covert speech in healthy participants allows to study the activity underlying speech production in the absence of sensory feedbacks, similarly to what could be observed in paralyzed people.

Our team had the opportunity to record the brain activity of a human participant in Broca's area, a key region for speech production, with a micro-electrode array. We used this data to analyze the difference between activities elicited by overt and covert speech and to explore the possibility of using spiking activity in Broca's area to drive a speech BCI. **Our second objective was therefore to gain novel insights the neuronal activities during overt and covert speech in Broca's area.**

Calibrating a BCI for participants unable to speak is especially challenging because, in the absence of behavioral output, it is not possible to precisely know the timing of inner speech production. Since most current decoding models are calibrated using simultaneously recorded overt speech and neural signals, new calibration strategies must

be developed. **Our third objective was to calibrate a speech decoder based on electrocorticography of the left hemisphere during covert speech.**

Organization of the manuscript

The first four chapters introduce the different topics covered in this thesis:

- **Chapter 1** presents the concept of brain-computer interfaces and the characteristics of existing interfaces regarding brain activity recording and feedback. It concludes with the current state of development of BCI for communication.
- **Chapter 2** exposes the different representations of speech and how these can be used to quantify and synthesize speech.
- **Chapter 3** gives an overview about the current knowledge on the functional organization of speech production in the human brain.
- **Chapter 4** describes the different aspects of speech decoding based on electrophysiological recordings and reviews the performance of important studies in this field.

The next three chapters develop the main results of the thesis:

- **Chapter 5** presents the phenomenon of acoustic contamination of electrophysiological signals. We show that it is a widespread problem worldwide that can possibly bias the results of speech decoding studies. We expose its probable cause and a way to objectively assess its presence in recordings.
- **Chapter 6** presents an analysis of the activity of a population of neurons in Broca's area showing different dynamics during overt and covert speech. We also show that there is a relationship between the phonological content of overtly pronounced sentences and the spiking patterns across the population.
- **Chapter 7** shows preliminary results concerning the continuous decoding of speech features during the covert production of phonemes. It exposes the offline decoding results based on the recording of one participant and describes an attempt to perform online speech decoding on the same subject.

Finally, the last chapter summarizes the main contributions of the thesis and discusses possible directions for future works.

Chapter 1

Brain-computer interfaces

This chapter introduces the concept of BCI. The different ways to record brain activity are exposed, as well as the interactions of the system with the user. Finally, the current state of BCIs for communication is summarized.

1.1 Principles

A BCI is a system that measures brain activity and converts it into commands for an external device. The interest of such system is to allow the user to interact with a device using a pathway that does not involve the conventional brain outputs provided by peripheral nerves and muscles. The processing chain of a BCI system is conventionally broken down into its main components as shown in Figure 1. The brain activity is recorded in real-time using a measurement device (see section 1.2). The recorded signal is then usually preprocessed and some relevant features are extracted from it. Finally, a translation algorithm decodes these features and converts it into adapted commands for an external device. The activation of the external by the generated commands provides a feedback to the user. This feedback makes it possible for the user to learn to control the external device via his brain activity.

1.2 Brain activity recording

BCIs fundamentally depend on brain activity recordings. Brain activity can be monitored using many different methods. The present section introduces these methods and discusses their potential use in BCIs. The criteria that will mainly be discussed are their temporal and spatial resolutions as well as their portability and invasiveness. These characteristics are summed up in Table 1. What is referred to as temporal resolution in the present section is the time scale at which significant variations can be observed in the recorded activity. It is therefore related to the intrinsic frequency of the measured signal rather than its sampling rate.

Brain activity measurements can be separated into two main categories: metabolic and electrophysiological signals. The activation of neurons require energy that is brought in the form of oxygen and glucose via the vascular system. During a cognitive task, the

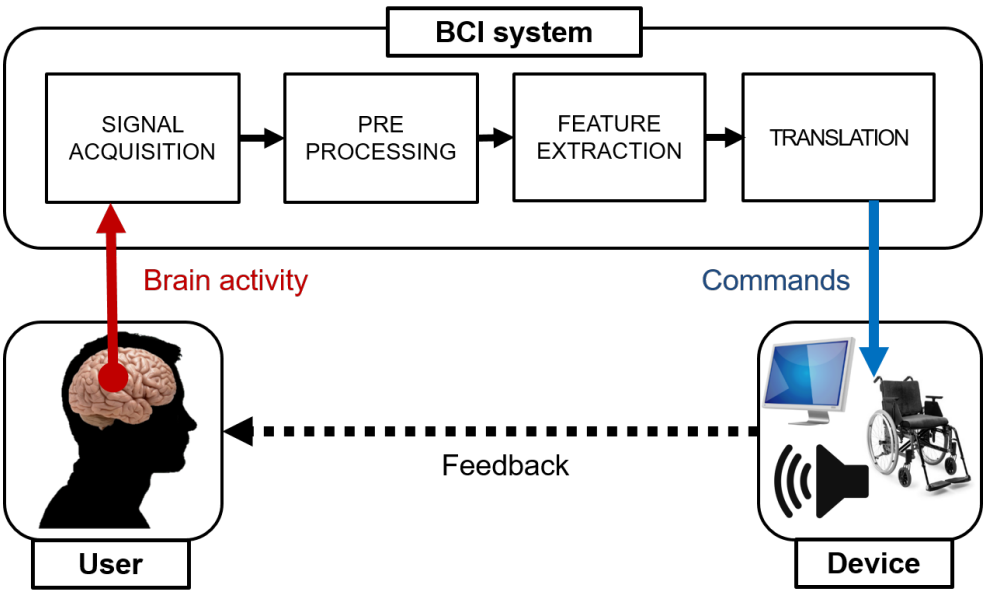


Figure 1. Block diagram of a BCI application.

blood flow adapts to meet the demand of brain areas depending on their involvement in the task. This phenomenon is called the haemodynamic response. Several methods can measure the variations of blood oxygenation – functional magnetic resonance imaging (fMRI), functional near infrared spectroscopy (fNIRS), optical imaging of intrinsic signals (OIIS) – or the changes in glucose concentration – positron emission tomography (PET). The spatial resolution of the haemodynamic measurement depends on the method, from a centimeter scale for fNIRS to an order of a hundred microns for OIIS. Metabolic signals however vary slowly and thus have a temporal resolution of about a second. A BCI based on metabolic activity can therefore only produce a very limited number of commands per unit of time.

Electrophysiological methods on the other hand measure the electrical signals resulting from the electrochemical activity of brain cells. Intracellular recordings offer a view of the electrical activity of single neurons. However, this method is not suitable for real-life BCIs as it cannot be carried out in freely moving subjects during long periods. The variations of electrical potential resulting from neuronal activity can also be measured using electrodes placed on the scalp, on the surface of the brain or inside the brain (see Figure 2). These signals are constituted by a combination of synaptic currents, action potentials and all the other ion exchanges through the membrane of neurons and glial cells. They are therefore composed of fast- and slow-varying components and can be measured with a millisecond temporal resolution. Contactless estimation of the internal currents can also be performed by measuring variations of the magnetic field at the level of small brain areas using magnetoencephalography (MEG). This technology has an interesting spatio-temporal resolution but requires a very complex equipment as well as a shielded room, which makes it difficult to envision as a tool for real-life compatible BCIs.

Electroencephalography (EEG) measures the differences of electrical potential on the

Method	Temporal resolution	Spatial resolution	Invasiveness	Portability
fMRI	∼1 s	∼1 mm	No	No
fNIRS	∼1 s	∼2 cm	No	Yes
OIIS	∼0.5 s	∼0.1 mm	Yes	Yes
PET	∼0.2 s	∼1 mm	No	No
EEG	∼10 ms	∼1 cm	No	Yes
MEG	∼10 ms	∼2-5 mm	No	No
ECoG	∼10 ms	∼0.5-5 mm	Yes	Yes
SEEG	∼10 ms	∼0.1 mm	Yes	Yes
MEA	∼1 ms	∼0.1 mm	Yes	Yes
Ca^{2+} imaging	∼10 ms	∼0.01 mm	Yes	Yes
VSD	∼1 ms	∼0.01 mm	Yes	Yes

Table 1. Comparison of the brain activity real-time measurement methods.

scalp. Due to the distance with the electrical sources and the low conductivity of the skull, signals coming from single neurons are distorted and averaged over large areas. It is therefore considered that EEG mostly reflects the activity of large groups of synchronous neurons. Millimeter-sized electrodes placed under the skull offer a more local measurement of the electrical potential, called local field potential (LFP), either on the surface of the cortex with electrocorticography (ECoG) or inside the brain with stereo-electroencephalography (SEEG). Micro-ECoG (μECoG) is used to designate ECoG grids with small-sized electrodes and short inter-electrode distances. These methods are invasive as they require the skull to be opened but they offer a much better spatial resolution by avoiding the distortion effects caused by bone. SEEG has the advantage of being less invasive than ECoG as is only requires very small openings in the skull for the implantation. Micro-electrode arrays (MEA) can have tens of electrodes either distributed at different depths or covering an area of a few square millimeters at constant depth. They have a very good spatiotemporal resolution but cover a very small region of the brain, which can be compensated by using several of them.

An alternative way to measure the activity of single cells is the calcium imaging which is an optical measurement of intra-cellular calcium ions (Ca^{2+}) concentration. This technique allows to resolve action potentials and to monitor tens of cells in a 3D volume. Its temporal resolution is limited by the intra-cellular calcium dynamics but processing methods have been proposed to reconstruct neuron action potentials (Yaksi and Friedrich, 2006; Grewe et al., 2010). Voltage-sensitive dyes (VSD) can also be used to simultaneously measure the firing activity of large group of neurons. The optical signal that this technique provides is linearly related to the membrane potential of the cells (Loew et al., 1985). Ca^{2+} and VSD imaging can be carried out with portable optical devices but requires to expose the cells to fluorescent dyes, which currently makes them incompatible with human recordings.

In conclusion, it appears that electrophysiological recordings are for the time being the most suitable measurements for real-life BCI applications due to their portability and

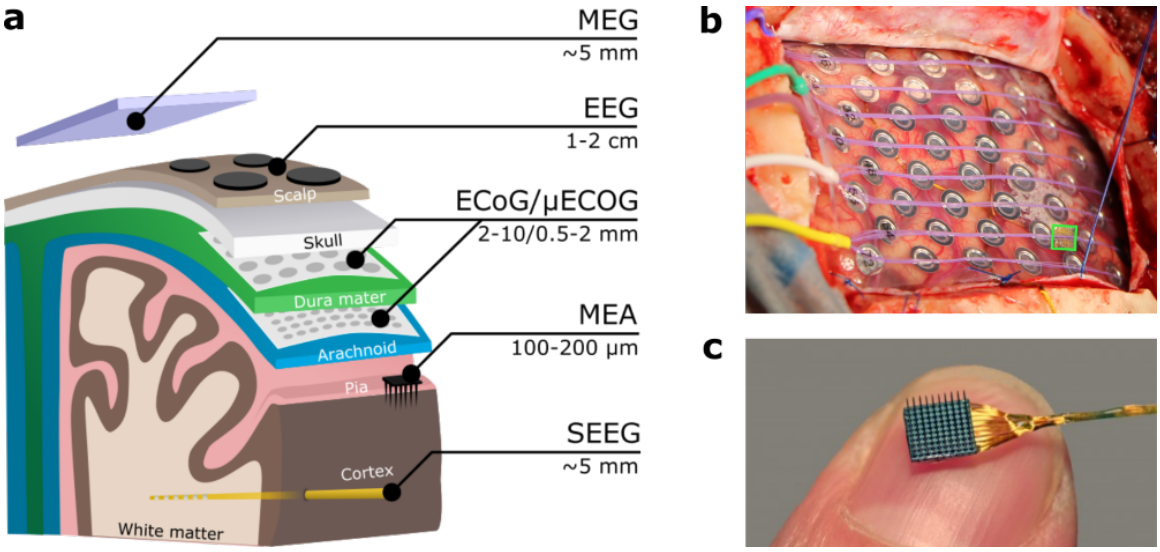


Figure 2. Electrophysiological measurement methods of brain activity. (a) Illustration of the the different measurement methods and the relevant biological tissues involved. Approximate spatial resolution are indicated. Image adapted from Jorfi et al. (2015) and Bocquelet (2017). **(b)** ECoG grid and a Utah array (green square) placed over the left hemisphere of a human subject. Photography taken during the implantation surgery of participant P5 (see section 5.2.1.2 for more details). (c) Photography of a Utah array. The Utah array is a type of rigid MEA with micro-electrodes located at the tip of silicon needles. This 4x4 mm array consists of 96 electrodes. Image credit: Stanford Neural Prosthetics Translational Laboratory.

high temporal resolution. Invasive probes in particular offer a good spatial resolution that allow to record modulation of brain activity in very specific regions. For these reasons, the most advanced BCI for motor prostheses have been based on invasive electrophysiology. In Wodlinger et al. (2015) and Benabid et al. (2019) for example, the participants were able to control a robotic arm based on intracortical micro-electrodes and an exoskeleton based on ECoG, respectively.

1.3 Feedback and adaptation

The desirable evolution of the interaction between the user and the BCI is generally described as a co-adaptive process (Wolpaw et al., 2000). Formally, depending on the type of adaptation, BCI systems can be divided into 3 categories (Sellers et al., 2007). These categories are represented in Figure 3. In the first case, the optimal BCI algorithm is seen as the solution of a conventional machine learning problem, where a model fits the brain activity of the user, seen as stable. Operant conditioning conceptualizes a BCI with a fixed algorithm, in which the progress is supposed to come from the adaptation of the user's brain activity – through modifications of the user's behavior and/or brain plasticity. Finally, the fact that a BCI algorithm takes into account the evolution of the user activity,

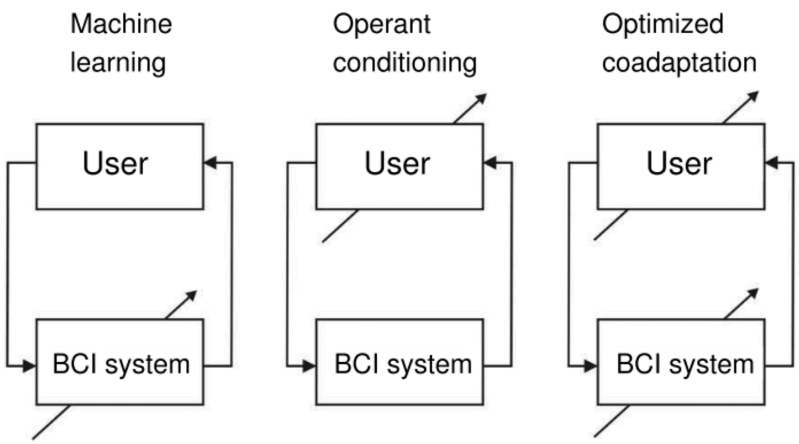


Figure 3. Categories of BCIs according to the adaptation paradigm. Reprinted from Sellers et al. (2007).

in particular as he learns to use the system, is referred to as optimized coadaptation.

In practice, the possibility for the user to adapt is dependent on the feedback that is provided by the system. Neuronal operant conditioning is perhaps the most direct example of the volitional modulation of brain activity. It was for instance demonstrated that non-human primates were able to increase and decrease the activity of neurons in the precentral cortex when provided with audiovisual feedback and food reinforcement Fetz (1969); Fetz and Baker (1973). More recently, (Moritz et al., 2008) showed that non-human primates were able to modulate the firing rate of motor neurons in the motor cortex to stimulate different muscles of their own artificially paralyzed arm. In humans, (Cerf et al., 2010) demonstrated that subjects can regulate firing rates of single neurons in the medial temporal lobe to make an image appear on a screen. It was also shown that operant conditioning can be used with other type of brain activity recordings. Indeed, subjects have been able to modify their slow cortical potentials when provided with an audiovisual feedback of their intensity (Elbert et al., 1980; Hinterberger et al., 2004) and to gain some control over their sensorimotor rhythms in the two hemispheres when provided with a real-time activity map of their motor cortex (Pineda et al., 2003).

Since users progressively learn to modulate their brain activities and because of the physiological and recording non-stationarities, a BCI algorithm that has been calibrated on a training dataset may lose some accuracy in prediction over time (Jarosiewicz et al., 2015). Such a calibration is described as open-loop, in opposition to closed-loop systems which incorporate a mechanism to update their prediction algorithm. Closed-loop calibration have been shown to obtain better results than open-loop when using the same amount of data (Gilja et al., 2012). They may allow for a better tuning of both the user's brain activity and the algorithm (Jarosiewicz et al., 2013).

1.4 Brain-computer interfaces for communication

Brain-computer interfaces have been used to provide a pathway for communication that is independent from muscular activity. To this end, BCIs were developed to allow users to spell words based on EEG signals. The P300 wave is an event-related evoked potential appearing when an expected event occurs in a sequence of other events. It was shown that it can be used to operate a spelling device (Farwell and Donchin, 1988), and was later successfully used by patients suffering from ALS to communicate (Nijboer et al., 2008). These systems have the huge advantage of being non-invasive. However, even in the latest developments of this technology it still takes users several minutes to type a single sentence (Townsend and Platsko, 2016). Operating these systems is demanding for the users who should be focused on the task, which limits their use for prolonged period of time (Käthner et al., 2014).

In parallel, the development of intracortical BCIs based on upper limb movement intention showed a progression over the years in terms of accuracy and of degrees of freedom (DoF). At first in animal studies, it was for example shown that rats could control a simple unidirectional robotic arm via the firing rates of selected neurons (Chapin et al., 1999). Later on, more complex control of 2D and 3D robotic arms via neuronal activity was achieved in studies involving non-human primates (Wessberg et al., 2000; Serruya et al., 2002). Intracortical BCIs in humans also allowed to achieve increasingly complex tasks: 2D cursor control (Hochberg et al., 2006), 3D reach and grasp tasks (Hochberg et al., 2012), control of robotic arms with 7 and 10 DoF (Collinger et al., 2013; Wodlinger et al., 2015). These BCIs, while having the major drawback of being invasive, have been reported to require less mental workload from the user (Hochberg et al., 2006), as the external device seem to become progressively embodied along the training. Increasing control accuracy made it possible to achieve better results than P300 spelling devices with virtual typing using a 2D cursor controlled by hand movement intention, reaching an average typing rates of about 30 correct characters per minute (Pandarinath et al., 2017).

If BCIs based on the activity of the upper limb motor cortex seem to enable users to have a more natural control, they still do not provide a communication pathway that is comparable to speech in its efficiency. This motivates the research for the development of a BCI that would enable to synthesize continuous speech, as pioneered by Guenther and colleagues (Guenther et al., 2009), who were able to synthesize vowels from the brain activity of a locked-in person. Such a BCI records brain activity in speech-specific areas, extracts informative features from it, decodes it into control parameters and feeds a speech synthesizer producing real-time audio feedback to the user.

Chapter 2

Quantifying and synthesizing speech

Speech is the main oral communication system used by humans. It conveys information as sounds that are the result of vocalization. As opposed to other forms of vocalizations, speaking consists in forming sequences of elementary signs, phonemes. Defined sequences of phonemes make up words, which constitute symbols in language systems. Building a speech BCI requires being able to decode speech features from brain activity and to re-synthesize an audio signal from it. In order to find such features, speech can be analyzed and quantified on various levels, from the way phonemes are pronounced to the meaning of words.

2.1 Physical characteristics

2.1.1 Phonemes

As introduced above, the words of a language are orally expressed by the production of a sequence of elementary sounds called phonemes. Phonemes refer to groups of speech sounds that are perceived to have the same characteristics by speakers of a particular language. The set of phonemes vary depending on the language or dialect. For a defined language, the number of phonemes also vary according to the interpretations of linguists. The approximate number of phonemes in Standard French, General American and Oxford English are considered to be 33∼39, 38∼40 and 44∼49, respectively. In the following, phonemes will be transcribed using the International Phonetic Alphabet (IPA; see chart in Annex 1).

The phonemes are distinguished based on their acoustical properties. Breaking down speech as a sequence is important to understand its structure and can be meaningful for the analysis of the neural processes underlying its production and perception. These sounds are shaped in the human vocal tract by a process called articulation. These physical characteristics of speech are described in the following sections.

2.1.2 Speech production mechanisms

The production of sounds in speech is physically performed by muscular effectors that are controlled voluntarily by the central nervous system. The sound is produced by the airflow

coming from the respiratory system and going through the vocal tract. The shape and movements of the vocal tract modulate this sound. The main organs of the vocal tract are represented on Figure 4.

2.1.2.1 Vocal cords

Coming out of the lungs, the airflow first goes through the larynx which houses the vocal cords (or folds). When brought together close enough, the vocal cords vibrate as the air passes through. The presence or absence of vibration of the vocal cords distinguishes voiced sounds from unvoiced sounds. For example, phoneme /g/ is the voiced equivalent of phoneme /k/ – both are produced with the same vocal tract configuration but are distinguished by the presence of vocal cords vibrations for the former.

2.1.2.2 Articulators

The acoustical properties of both voiced and unvoiced sounds are modulated by the movement of key organs called articulators. The positions of key articulators, namely the lips, tongue, jaw and velum, mainly influence the shape of the vocal tract and therefore the produced sounds.

Phonemes are divided into vowels and consonants depending on the way they are produced. Vowels are produced without any stricture of the vocal tract (like /a/, /i/, /u/) while consonants are produced with complete or partial closure of the vocal tract (like /t/, /m/, /f/). In most languages, including French and English, vowels are voiced.

Consonants can be grouped into different categories depending on the place of articulation. The place of articulation refers to the location of the contact point where the obstruction of the vocal tract occurs. Consonants whose place of articulation is the lips (like /m/), the upper teeth (like /θ/) or the velum (like /k/) are for example called labial, dental or velar consonants, respectively.

Phonemes are also characterized by the involvement of the nasal cavity in the production of the sound. The nasalization of a sound can be considered as a binary feature, determined by the position of the velum during production. During speech, the velum can retract and elevate to separate the nasal cavity from the oral cavity. When this separation is incomplete, some or all of the airflow passes through the nasal cavity, modifying the acoustic properties of the produced sound. In English, consonants /m/ and /n/ are the only nasal sounds. In French, nasal vowels like /ɑ̃/ and /ɔ̃/ also exist.

Despite the fact that speech consists in a sequence of phonemes, articulation underlying speech remains a continuous process. Thus, the position of articulators during the production of a given phoneme is influence by the context. Influences of preceding and following phonemes are respectively called anticipatory and carryover coarticulation. These phenomena have to be taken into account when considering speech recognition or synthesis.

2.1.2.3 Measuring articulatory movements

The movements of the articulators can be measured using different methods. The shape of the vocal tract can be fully monitored using X-ray imaging and MRI. These techniques have

a good spatial and temporal resolutions but require large equipments that forces subjects to lie down in the case of MRI. Exposition to X-ray should also be limited for safety reasons. Other recording methods, such as video recording and tongue ultrasonography allow to get partial information about the articulators (Hueber et al., 2010). Lips and jaw positions can be extracted from the video while ultrasonography allows to measure the movements of the tongue. These two methods have the advantage of being compatible with simultaneous brain electrophysiological recording, which make them attractive for the study of speech production (Bouchard et al., 2016). Finally, electromagnetic articulography (EMA) allows three-dimensional tracking of small sensor coils placed near a magnetic field generator, with high spatial (< 1 mm) and temporal (up to 400 frames/s) resolutions. This recording system necessitates to attach wired sensors on the different articulators which can be uncomfortable but can be used to monitor the position of the most relevant points of the vocal tract.

2.1.2.4 Inferring articulation from sound

Specific phonemes tend to be associated with specific articulatory gestures. Since it is not simple to precisely record articulation, techniques have been developed to estimate articulatory positions from corresponding audio data. However, it should be noted that different positions of the articulators can produce the same sound. For instance, the precise position of the tongue is not relevant to the production of the labial consonants (like $/\mathrm{m}/$ or $/\mathrm{b}/$). Many methods have been proposed to solve the ill-defined problem of articulatory-to-acoustics inversion, including Gaussian mixture models (GMMs), hidden-Markov models (HMMs) and artificial neural networks (ANNs; see Illa et al., 2017 for review). In particular, speaker-independent acoustic-to-articulatory inversion based on ANN has been recently used to study the neural correlates of articulatory trajectories in ECoG recordings (Chartier et al., 2018; Anumanchipalli et al., 2019).

2.1.3 Acoustics

2.1.3.1 Spectro-temporal structure

In French and in English, phonemes last from about 30 ms to 400 ms (House, 1961; Umeda, 1977; O'Shaughnessy, 1981) with consonants being generally shorter than vowels. Phonemes have different spectral characteristics depending on the way that they are produced. During voiced phonemes, the vibration of the vocal cords generates a periodic sound. The periodicity of the voiced sounds is visible in the frequency domain by the presence of a fundamental frequency and harmonics. The fundamental frequency of voiced sounds determines the pitch of the speaker's voice.

Vowels are voiced sounds characterized by a stable frequency structure. When producing vowels, the vocal tract acts as a static acoustic filter modulating the amplitude of the different harmonics of the periodic sound generated by the vocal cords. The frequencies corresponding to maxima in the spectrum of the resulting sound are called formants. The formant with the lowest frequency is called F_1 and the one with the second lowest frequency is called F_2 (see Figure 6a). The vowels can most often be distinguished from

one another by looking at the frequencies of the first two formants (see Figure 6b).

During voiced consonants such as /g/ or /m/, the vocal folds also produce a locally periodic sound. However, because the transient nature of articulatory gestures involved, the fundamental frequency of this sound changes over the time course of the phoneme. These sounds are therefore characterized by both the values of the formants and their variations.

Unvoiced consonants correspond to non periodic sounds and the associated spectral power is therefore spread on large frequency bands. The different unvoiced consonants can still be distinguished based on the spectral power distribution. This distribution can be stable (like for /s/ or /ʃ/) or vary over the time course of the phoneme (like for /t/ or /k/).

2.1.3.2 Acoustic representations

Given the spectro-temporal structure of speech sounds, spectrograms constitute an adapted representation which allow for example to easily distinguish voiced sound from others. Specific representations of speech sounds have also been developed in order to capture the characteristics of speech sounds in a more compact way. The present sections presents three of them: the auditory spectrogram, the mel-frequency cepstral coefficients and the mel-cepstral filter coefficients.

Representations have been obtained by mimicking the processing steps that are thought to occur in the human auditory system. The method proposed by Yang et al. (1992) to compute these so-called auditory spectrograms is summarized by the following operations:

1. Computing a spectrogram-like representation using a cochlear filter bank
2. Low-pass filtering and applying a non-linear compression to imitate hair cells' behavior
3. Computing the first order derivative along the frequency axis and rectifying it to approximate the effect of the lateral inhibitory network
4. Finally, integrating the result to simulate the patterns observed on the auditory nerve

This representation can effectively describe speech sounds with 128 spectral coefficients and can be used to reconstruct the audio signal using an iterative convex optimization algorithm (Chi et al., 2005).

The vocal tract can be viewed as a filter modulating a source sound produced by the airflow. The effect of the filter can be inferred from the spectral envelope of the signal. Relevant characteristics of this envelope can be extracted using cepstral analysis. In particular, mel-frequency cepstral coefficients (MFCCs) are a commonly used representation of speech sounds in applications such as speech recognition. MFCCs are computed as follow:

1. Computing the Fourier transform of the speech signal
2. Mapping the obtained power spectrum onto the mel scale using triangular overlapping windows

3. Computing the logarithm of the mel-frequency spectrum
4. Applying the discrete cosine transform to the mel-frequency log spectrum

The MFCCs are the amplitudes of the resulting cepstrum.

A similar representation uses mel-cepstral coefficients that are explicitly defined as the coefficients of a filter, the so-called Mel Log Spectrum Approximation (MLSA) filter (Imai, 1983). These coefficients are simply denoted as MELs in the present manuscript. In this method, the spectral envelope of the sound is modeled by the vocal filter H composed by M MELs $c_\alpha(m)$ and defined as follow:

$$H(z) = exp \sum_{m=1}^{M} c_\alpha(m) \cdot \tilde{z}^{-m}$$

$$\text{with } \tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}$$

The coefficient α is set so that the mel-scale becomes a good approximation of the human sensitivity to the loudness of speech. In order to reconstruct the corresponding speech, the obtained filter can be excited with a source signal representing the vocal folds activity – generally a train of pulses for voice sounds and white noise for unvoiced sounds.

2.2 Semantic representations of speech

Speech can also be analyzed as a sequence of words. Unlike phonemes which constitute an ensemble with a limited number of elements (less than 50 for French and English), the vocabulary used in everyday conversations comprises thousands of words. Words being composed by a sequence of phonemes, they can be represented as acoustic or articulatory trajectories. They can also be analyzed according to their symbolic function in language, i.e their meaning.

The meaning carried by words can be analyzed and categorized in different ways. In particular, it is possible to project them in finite-dimensional vector spaces, called word embeddings, in which distances reflect the semantic similarities between words. Distributional semantics is a discipline that aims at constructing such semantic representations of words based on the way they are distributed in large samples of data. These representations are built on the hypothesis that words occurring in the same contexts have similar meanings (Harris, 1954). This quantification of words' similarities has proven useful in applications of computational linguistics such as machine translation. A recent pre-print shows that these vectorized representation of words might be interesting target features to decode speech from neural activity (Affolter et al., 2020).

Since the 2010s, DNN-based models such as word2vec (Mikolov et al., 2013) have outperformed the other distributed models. In word2vec model, a DNN is trained to predict a word from its context (continuous bag-of-words variant) or to predict context words from a single word input (skip-gram variant). In both variants the vectorized representation of words is produced by a hidden layer of neurons whose size reflects the dimensionality of the resulting word vectors. Mikolov and colleagues report that the skip-

gram variant is better at capturing the words' semantic relationships than the continuous bag-of-words variant.

2.3 Speech synthesis

Modern computer-based speech synthesis can be achieved using different techniques that can be divided in three main categories: formant synthesis, articulatory synthesis and text-to-speech synthesis.

2.3.1 Acoustic synthesis

Acoustic synthesis (often called formant synthesis) uses input parameters that describe the acoustic content of the target sound. The synthesized speech is obtained by modulating an excitation signal, which represents the activity of the vocal cords, using a time-varying filter. The excitation signal mainly differentiates voiced and unvoiced sounds but can integrate more information depending on the models. Similarly, the target acoustic parameters used to construct the filter vary. Speech synthesizers have for example been based on the frequencies of the first formants (Klatt, 1980) or on mel-cepstral parameters (Imai, 1983; Kawahara, 1997; Morise et al., 2016).

2.3.2 Articulatory synthesis

Articulatory synthesis refers to techniques that synthesize speech by simulating the physical process of speech production. In that case, the input parameters are trajectories of articulatory features – for example the positions over time of the main speech organs such as the tongue, the jaw, the lips or the velum. Two types of approaches have been developed to this end: physical approaches that model the geometry of oral cavities and their acoustic properties, and non-physical approaches that exploit large articulatory-acoustic databases with machine learning techniques to model the relationship between the two representations. In both cases, a preliminary step consists in acquiring articulatory data, using the techniques described in section 2.1.2.3.

The physical approaches are based on models that simulate the geometry of the vocal tract during articulation. Two-dimensional models of the oral cavities in the midsagittal plane were first proposed (Perkell, 1974; Maeda, 1990). Later on, three-dimensional models were developed to overcome shortcomings of the former ones (Engwall, 1999). To generate the sound waveforms, simulations of the sound propagation can be applied to these approximate geometries. 3D models of propagation yield better results but require finite elements simulations with heavy computations. On the other hand, 1D models simulating plane waves in a 2D geometry are less precise but can run in real time.

The non-physical approaches rely on supervised machine-learning methods in order to capture the relationship between simultaneously recorded articulatory and acoustic data. These so-called articulatory-to-acoustic mappings can be divided into two categories: those that first use articulatory speech recognition and those that directly estimate the

trajectories of the acoustic parameters. In both cases, an acoustic synthesizer is then used to convert the acoustic trajectories into speech waveforms.

The first type of non-physical approaches transforms articulatory positions into a discrete intermediate representation (such as a sequence of phones or words) which is then fed to an appropriate synthesizer (see section 2.3.3 about text-to-speech synthesis). Building the intermediate representation requires to carry out speech recognition on articulatory data, which showed promising results but only when using reduced sets of words or phonemes (Jou et al., 2006; Hofe et al., 2013).

The second type of non-physical approaches uses statistical models – such as artificial neural networks (ANNs; Kello and Plaut, 2004), Gaussian mixture models (GMMs; Toda et al., 2008) or hidden Markov models (HMMs; Hueber et al., 2012) – to achieve the articulatory-to-acoustic transformation. The most advanced results were obtained using deep neural networks (DNNs) to map EMA data to acoustic features. For example, a word recognition accuracy over 95% was obtained in Bocquelet et al. (2016c). A DNN was also successfully used by Anumanchipalli et al. (2019) to synthesize intelligible speech using articulatory trajectories that had been decoded from brain activity.

2.3.3 Text-to-speech synthesis

Finally, the most common category of speech synthesis is text-to-speech synthesis, for which the input is typically a sequence of words. It is usually composed of two modules. The first one converts the text into a sequence of phonetic units (such as phonemes or syllables) along with additional information that influences the target sound (like the surrounding phonemes). The second module generates a waveform based on the produced sequence of phonetic units. This second step can be achieved using two types of methods: the concatenative synthesis and the statistical parametric synthesis. In concatenative synthesis, pre-recorded audio speech segments corresponding to the phonetic units are concatenated. In statistical parametric synthesis, a model is used to convert the sequence of phonetic units directly into the speech signal or a parametric representation of it, such as the ones used in formant synthesis. In recent years, previously used methods such as HMMs have been outperformed by deep neural networks (Ze et al., 2013). In 2016, artificial neural networks were even used to directly predict the speech waveform (without specifying an intermediate representation) and obtained better subjective evaluations than concatenative models (Van Den Oord et al., 2016).

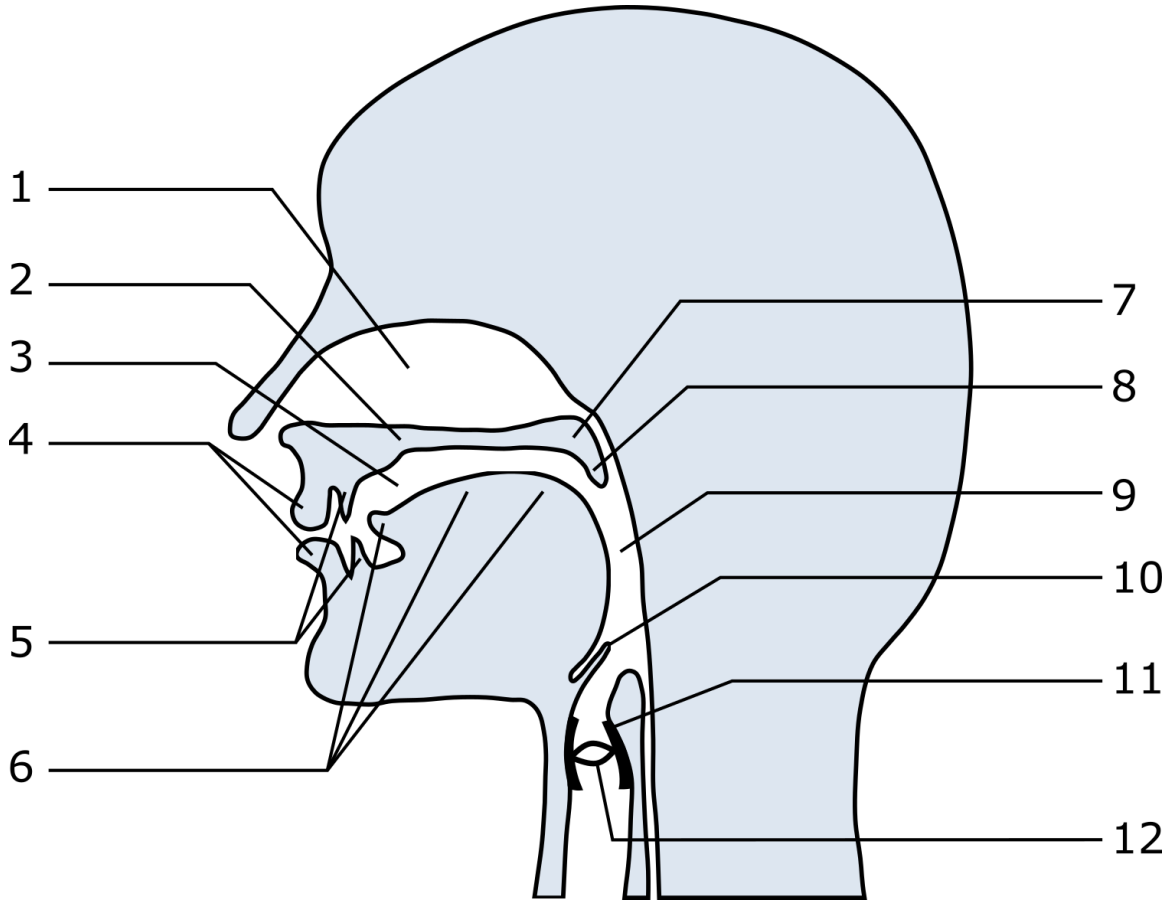


Figure 4. Illustration of the vocal tract with its main organs and landmarks. The nasal cavity (1) is a set of connecting chambers in the front part of the face. It is separated from the oral cavity (3) by the the hard palate (2). The opening of the oral cavity and its shape are controlled by the positions of the teeth (4) and the lips (5). The tongue (6), which can be decomposed into its front, middle and back parts, affect the shape of the oral cavity. The soft palate or velum (7) is connected to the hard palate and can move to close off the airway between the oral and nasal cavities. The uvula (8) hangs from the velum. The pharynx (9) connects the mouth to the trachea and the esophagus. The epiglottis (10) closes the trachea during swallowing. The vocal cords (11) or folds are located inside the larynx (12).

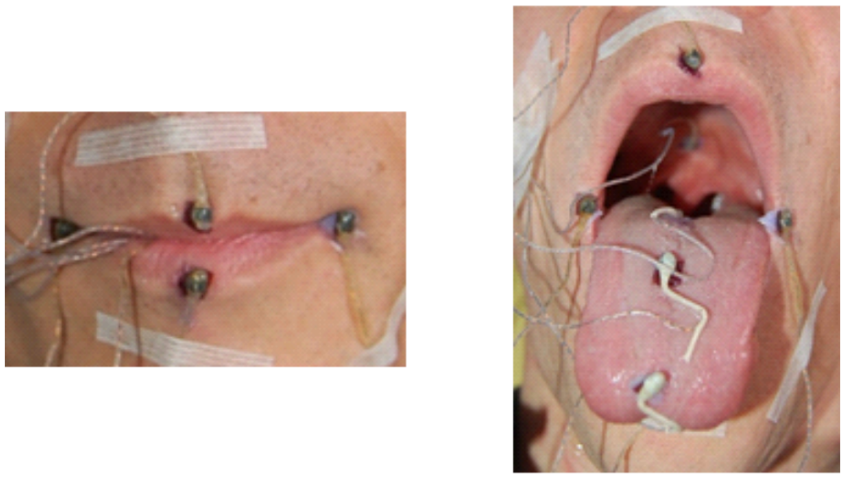


Figure 5. Electromagnetic articulography (EMA). Coils are glued to the lips, the jaw, the tongue and velum. The jaw sensor was glued at the base of the incisors and is not visible the images. This set-up was used for the recording of the BY2014 articulatory-acoustic corpus presented in section 7.2.3. Reprinted from Bocquelet et al. (2016c).

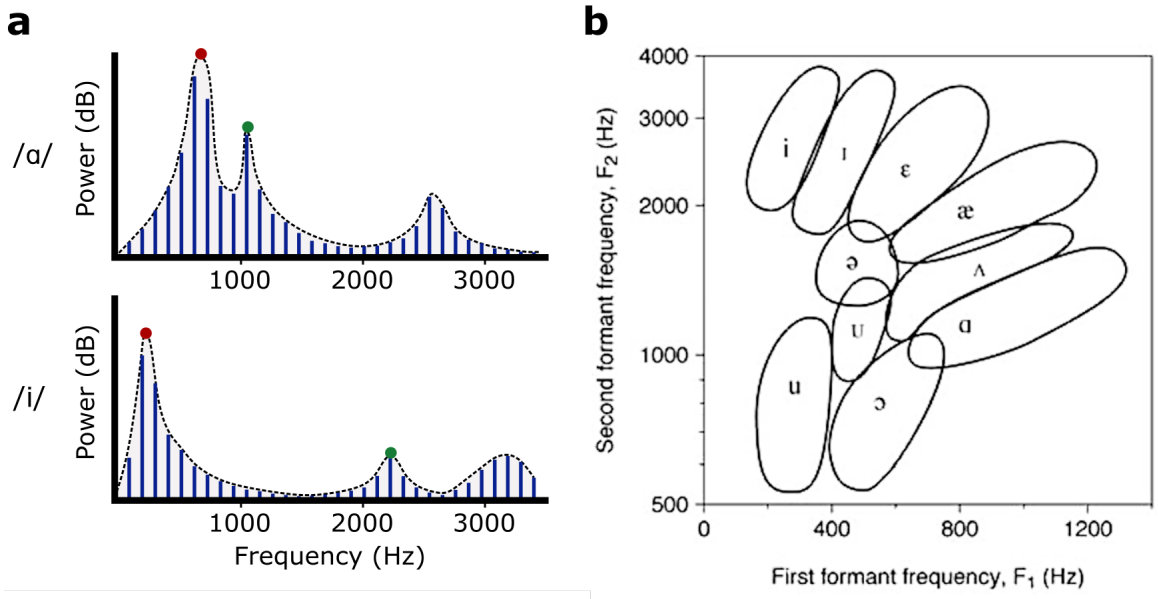


Figure 6. Speech formants. (a) Illustration representing typical spectral power distributions for vowels /ɑ/ and /i/. The power of the fundamental frequency and its harmonics are represented as blue bars. The spectral envelope is displayed as a dotted line. The red and green dots indicate the first and second formants, respectively. **(b)** Approximate distribution of the first and second formants of 10 English vowels. The distributions were estimated from the data of 76 speakers. Graph adapted from Peterson and Barney (1952).

Chapter 3

Neural basis of speech production

Understanding the functional organization of speech production in the brain is important for the development of speech BCIs as it guides the choices of the recording location and of the decoded speech representation. This chapter summarizes the current knowledge and hypotheses about the regions involved in speech production. First, an overview of the neuroanatomy of speech production is given through the dual-stream model of speech processing and the supposed roles of the inferior frontal gyrus are detailed. Then, models of speech motor control are discussed. Finally, the case of covert speech is exposed.

3.1 Cortical organization of speech production

3.1.1 The dual-stream model of speech processing

Current findings indicate that speech is the result of a highly interactive process between different brain structures specialized in the processing of specific speech representations or playing the role of an interface between representations. The most frequent description of the speech network is the dual-stream model of speech processing (Hickok and Poeppel, 2000, 2004, 2007). This model decomposes the language network into a ventral information stream linking auditory and semantic representations and a dorsal stream linking auditory and articulatory representations. This distinction between ventral and dorsal regions has been shown to be relevant when trying to delimit the anatomical substrates of speech comprehension and production, using lesion studies (Fridriksson et al., 2016) or tractography (Saura et al., 2008). The speech network proposed by Hickok and Poeppel (2007) is summarized in Figure 7. The speech-related cognitive functions have been reported to be strongly lateralized, with a dominance of the left hemisphere for the majority of the population (for 95–99% of right-handed individuals and about 70% of left-handed individuals according to Corballis, 2014). In the most common case of left lateralization, represented on Figure in Figure 7, the dorsal pathway in particular is highly left hemisphere dominant.

According to the dual-stream model, a spectrotemporal representation of speech is found in the dorsal part of the superior temporal gyrus (STG). This claim is coherent with Mesgarani et al. (2014) showing that local activity in the STG is tuned to phonetic features

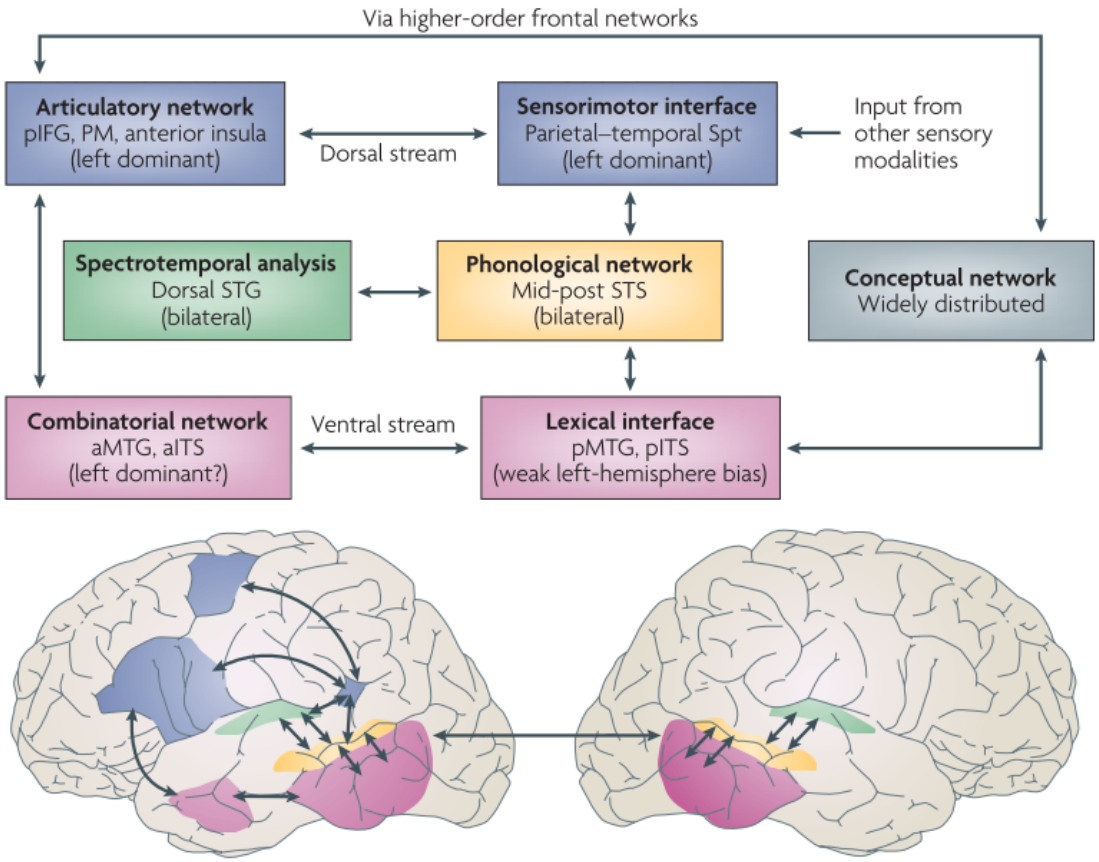


Figure 7. Schematic diagram of the dual-stream model of speech and approximate anatomical location of its components. The dorsal and ventral pathways are represented in blue and red, respectively. STS, superior temporal sulcus; STG, superior temporal gyrus; Spt, Sylvian parietal temporal region; aITS, anterior inferior temporal sulcus; aMTG, anterior middle temporal gyrus; MTG, anterior middle temporal gyrus; pIFG, posterior inferior frontal gyrus; PM, premotor cortex. Reprinted from Hickok and Poeppel (2007).

rather than single phonemes and Hullett et al. (2016) finding a topographic distribution of spectrotemporal modulation tuning in the same area. While these organizations were revealed in speech perception studies, this area is clearly active during speech production (Hickok and Poeppel, 2000; Tourville et al., 2019). This can be explained by the simple perception of one's own voice or by the presence of acoustic feedback in the motor control (see section 3.2).

The low-level auditory features found in the STG would then be associated to their phonological representation in areas surrounding the superior temporal sulcus (STS). Using several studies, Hickok and Poeppel (2007) indeed reveal sites in the STS that show specific activation to the processing of phonemic information.

On the one hand, this phonological network is connected to the ventral pathway, composed of areas in the posterior and anterior regions of the temporal gyrus. The

posterior areas are thought to perform lexical association while the anterior areas would be more specialized in syntactic and compositional semantic operations. There is also evidence that the ventral pathway also extends from the anterior temporal lobe to *pars orbitalis* (see Figure 8 and section 3.1.2.1), an anterior region of the inferior frontal gyrus (IFG; Saura et al., 2008; Berezutskaya et al., 2017).

On the other hand, the dorsal pathway connects the phonological network with an area in the Sylvian fissure at the parieto-temporal boundary (Spt), which is thought to act as an interface between auditory and motor representations of speech. This sensorimotor interface is thought to be linked to the premotor cortex (PM), the anterior insula and regions of the IFG, including the region historically known as Broca's area. These regions of the IFG are critically involved in speech production but it remains difficult to define their precise functions. The IFG appears to be involved in the translation from phonological to articulatory representation but also to have roles in manipulating the articulatory representation, as developed in section 3.1.2.1. As for the insula, it is thought to contribute to the motor planning or the motor control of articulatory movements (Dronkers, 1996; Ackermann and Riecker, 2004). Finally, the ventral part of the sensorimotor cortex displays a somatotopic organization corresponding to the orofacial effectors responsible for articulation, as discussed in section 3.1.2.2.

3.1.2 Roles of frontal areas in speech production

3.1.2.1 Inferior frontal gyrus

The IFG is commonly divided into three anatomical regions: *pars orbitalis*, *pars triangularis* and *pars opercularis* (see Fig 8). According the most common anatomical definition, Broca's area (BA) corresponds to the region formed by *pars triangularis* and *pars opercularis* (Tremblay and Dick, 2016). However, other definitions exist and some claim that Broadmann's cytoarchitectonic subdivision might not be precise enough to describe the complexity of this area (Amunts et al., 2010). This region is known to be critically involved in speech production but its precise roles remain unclear.

In the dual-model of speech processing, BA is identified as a part of the articulatory network (Figure 7. The essential role of BA in articulation is supported by lesion studies, Trupe et al. (2013) for example established a link between lesions in BA and speech motor planning disorders. However, the fact that articulatory representations are processed in BA remains a debated issue. Indeed, based on the direct cortical stimulation of 165 patients, Tate et al. (2014) concluded that BA seemed to be "modulating higher aspects of language such as semantic and phonological contents". High level functions have been attributed to BA or its subregions in several other studies. In their meta-analysis of imaging studies, Indefrey and Levelt (2004) proposed that the IFG is responsible for the syllabification, i.e. the compilation of a word's phonological representation into a sequence of syllables. This sequence would then be forwarded to the motor regions and translated there into articulatory code.

In line with this hypothesis, several studies claim that the major role of BA in speech production is to act as an intermediary between the acoustic and articulatory representation of speech. In the DIVA model of speech production (exposed in section 3.2), the left

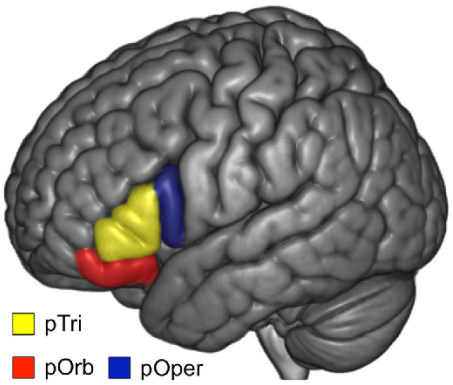


Figure 8. Location of the inferior central gyrus and its subregions. The inferior frontal gyrus (IFG) is divided into three standard anatomical subregions. pOper, *pars opercularis*; pTri, *pars triangularis*; pOrb, *pars orbitalis*. Reprinted from Matchin (2018).

posterior IFG and the ventral primary motor cortex are thought to hold a "speech sound map" that links auditory and articulatory representation of speech sounds (Guenther and Vladusich, 2012). This hypothesis that BA acts as an intermediary between the acoustics and articulation is supported by the study of ECoG recordings by Flinker et al. (2015). Their study highlighted the temporal dynamics of BA in a word repetition task, showing that information seemed to be sequentially passed on from the STG to BA during listening, from BA to the motor and auditory cortices during the pre-articulation phase and back from the motor cortex to BA during articulation. Mugler et al. (2018) also showed that it was possible to decode articulatory gestures and phonemes from the high-gamma activity in the IFG, suggesting that it would deal with both representations.

Finally, different high-level roles have been attributed to BA in imaging studies about speech processing. Sahin et al. (2009) showed that BA seems to be involved with phonological, lexical and grammatical processing. Meta-analyses of fMRI studies showed that a dorsal part of *pars opercularis* seems to be specialized in syntactic processing (Vigneau et al., 2006; Price, 2010). Price (2010) attributed a role of word selection to *pars triangularis* and a role of articulatory planning to the ventral part of *pars opercularis*. Hagoort and Indefrey (2014) state that the dorsal and ventral regions of the IFG seem to be specialized in syntactic and semantic processing respectively and Goucha and Friederici (2015) found the same functional distinction for *pars opercularis* and *pars triangularis*. The role in syntactic processing attributed to BA would however be due to a confusion to its function in short-term phonological memory, according to Rogalsky and Hickok (2011).

3.1.2.2 Ventral sensorimotor cortex

Originally using electrical stimulation (Foerster and Penfield, 1930) and recently using ECoG recordings (Bouchard et al., 2013; Chartier et al., 2018) and fMRI (Grabski et al., 2012; Carey et al., 2017), it has been shown that the ventral part of the sensorimotor cortex (vSMC) exhibits a rather consistent somatotopic organization of the speech articulators. Early stimulation studies showed that even though motor responses were

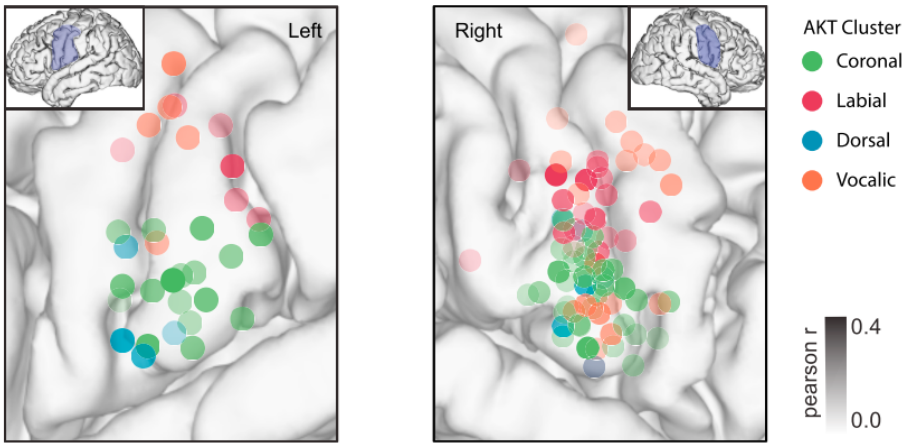


Figure 9. ECoG electrodes from five participants colored by type of vocal tract gesture mapped on common MRI-reconstructed brain. Opacity of electrode varies with Pearson's correlation coefficient from an encoding model of articulatory kinematic trajectory (AKT). This result shows that there is an overall subject-independent somatotopic organization of the articulatory gestures on the vSMC. Coronal and dorsal gestures refer to the front and back of the tongue. Vocalic gesture refers to voicing activity. Reprinted from Chartier et al. (2018).

most often elicited over the precentral gyrus and sensory responses on the postcentral gyrus, both types of responses were observed on both sides (Penfield and Boldrey, 1937). Chartier et al. (2018) derived a somatomotor map of vocal tract gestures from ECoG recordings that is coherent with these claims (see Fig 9). In the study by Mugler et al. (2014), it was also shown that articulatory gestures could be classified using high-gamma activity of the vSMC better than phonemes. Interestingly, Cheung et al. (2016) found that the activations of the vSMC were organized according to somatotopic representations of articulators during speech production but were distributed according to acoustic features during listening.

3.2 Models of speech motor control

Conceptual and computational models have been used to further understand the processes at stake in speech production, in particular the involvement of auditory processes in the motor control of the articulators. Computational models have been proposed to explain the auditory-motor interaction in the framework of feedback control. This idea that the online control of speech production relies on sensory feedback was first proposed by (Fairbanks, 1954) and formed the basis of later computational models, such as the DIVA model of speech production and acquisition (Directions Into Velocities of Articulators; Guenther et al., 1998). However, in order to be stable, feedback control requires nonnoisy, undelayed feedback, which is unrealistic for real sensory feedback according to Hickok et al. (2011). To address this issue, more recent versions of the DIVA model suppose the presence of

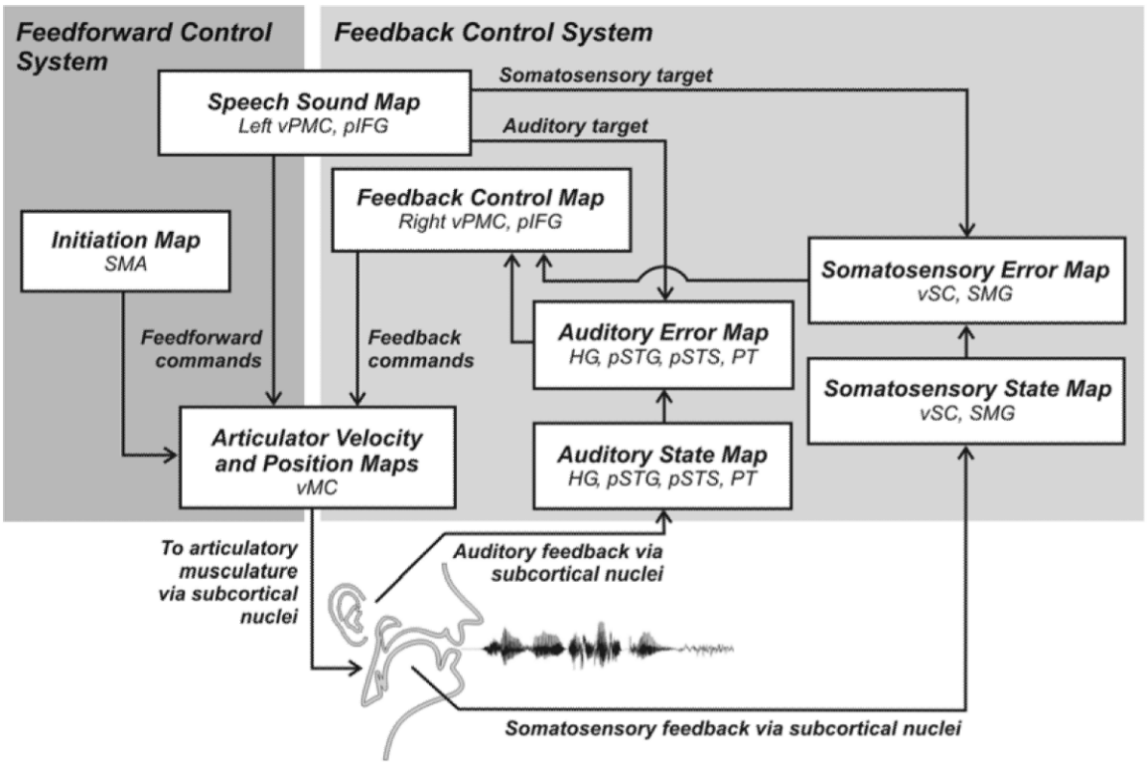


Figure 10. Schematic representation of the cortical components of the DIVA model of speech production and acquisition. In this model, the motor control of the articulators is achieved by the combination of feedback and feedforward control systems. Broca's area, situated in left posterior inferior frontal gyrus (pIFG), and the adjoining ventral premotor cortex (vPMC) constitute the interface between auditory and motor representations. SMA, supplementary motor area; vMC, ventral motor cortex; Heschl's gyrus; pSTG, posterior superior temporal gyrus; pSTS, posterior superior temporal sulcus; PT, planum temporale; vSC, ventral somatosensory cortex; SMG, supramarginal gyrus. Reprinted from Guenther and Vladusich (2012).

a feedforward controller (Guenther et al., 2006; Guenther and Vladusich, 2012). The different operations underlying the DIVA model have been associated to cortical areas, as can be seen on Figure 10. The plausibility of this model is supported by its ability to explain different experimental phenomena such as anticipatory/carryover coarticulation, as well as to successfully simulate spatial patterns of neural activity observed during speech production.

The presence of a feedforward controller has been formalized using the framework of state feedback control (SFC; Ventura et al., 2009). In the SFC hypothesis, an internal model receives an efference copy of the motor command and predicts the likely sensory feedback. The comparison between the target and the predicted feedback is then used to adapt the motor command with very low delay. The presence of SFC is supported by observations of cortical activity during speech production. For example the activation of premotor and auditory regions during imagined speech suggest the generation of sensory

estimates (Hickok et al., 2011). Houde and Chang (2015) also explain that the generation of an expected auditory response is coherent with phenomena of speaker-induced suppression and speech perturbation response enhancement. In both of these phenomena, it is observed that the mismatch between expected and actual feedback correlates with an increased activity in the auditory cortex.

In Hickok et al. (2011), the notion of SFC is extended to the integrated SFC model of speech production. In the integrated SFC model, three subsystems of the internal model are described: the motor phonological system, the auditory-motor translation and the auditory phonological system. First, the command to the articulators issued by the primary motor cortex are associated with a motor phonological representation in the premotor cortex. Then, this representation is translated via the Spt area to an auditory phonological representation in the auditory cortex. In the hierarchical SFC, Hickok (2012) proposes an extension of the integrated model that distinguishes two levels of internal forward models. The higher level (composed of the same areas as the integrated SFC) processes syllable representations while the lower level processes articulatory representations. In this model, the author proposes that the word production process goes from the word selection areas through the two types of internal models, which are then activated without resorting to an efference copy.

3.3 Covert speech

In opposition to overt speech, during which sounds are pronounced out loud, covert speech is an inner phenomenon, from one's mind to itself. Sometimes referred to as imagined speech, it can be considered as a mental simulation of speech. It is thought that this simulation partially reuses the neural processes of overt speech without producing an actual motor command. Psycholinguistic experiments show that covert speech can exhibit the same properties as overt speech – like varying in pitch and rhythm or containing phonological errors (see Perrone-Bertolotti et al., 2014 for review). The activation of articulatory representations during covert speech remains however a debated issue. Some researchers suggest that there are different subtypes of covert speech: an abstract covert speech using high level linguistic representations and a concrete covert speech characterized as either phonological or phonetic (Geva et al., 2011). According to Fernyhough (2004), different levels of internalization of speech might arise during development and be maintained in adulthood: 1) overt speech used in external dialogue, 2) subvocalized private speech used to speak to oneself, 3) expanded inner speech that is fully internalized but maintains the other properties of overt speech, 4) condensed inner speech that has lost the acoustical and structural properties of overt speech.

In agreement with the idea of continuum between overt and covert speech, many studies have shown that both types of task activate common brain areas. Martin et al. (2014) attempted to decode acoustic features from ECoG data during covert speech by using a model trained with overt speech data. The model obtained results over chance level and (Bocquelet et al., 2016a) showed a similar example in which a binary classifier trained on overt speech ECoG recordings in frontal areas also detected covert speech. These findings suggest that the neural correlates observed during overt and covert speech

share some similarities.

However, the activation patterns found in overt and covert speech differ to some extent. Some fMRI (Palmer et al., 2001) and ECoG (Pei et al., 2011b) studies have found greater activation in motor and premotor regions during overt speech, which is coherent with the absence of motor output and somatosensory feedback in covert speech. Evidence supporting the implication of motor areas in covert speech was however shown by Aziz-Zadeh et al. (2005): it was found that repetitive transcranial magnetic stimulation (rTMS) of left hemisphere frontal sites perturbed covert speech, resulting in longer latencies to perform a syllable counting task. Moreover, Guenther et al. (2009) used fMRI to determine the most active site during covert speech in a locked-in patient and found it to be a single area on the left precentral gyrus, lying near the border between premotor and primary motor cortex. Similarly, auditory areas have been found to be less activated during covert speech (Palmer et al., 2001; Pei et al., 2011b), which can be imputed to the absence of auditory feedback. The degree of activation of the auditory cortex might change depending on the fact that the subject is asked to perform kinesthetic or auditory imagery, as supported by the study of Martin et al., 2016. Other regions involved in overt speech have been reported to be active to a lesser extent during covert speech, in particular the left IFG, the left insula, the supramarginal gyrus, and the anterior cingulate (Perrone-Bertolotti et al., 2014).

Conversely, Perrone-Bertolotti et al. (2014) listed several areas that showed greater activation during covert speech in previous studies: the left precentral gyrus, left middle frontal gyrus, left or right middle temporal gyrus, left superior frontal gyrus, right cingulate gyrus, left or right inferior parietal lobe, left dorsal frontal cortex, left parahippocampal gyrus, right cerebellum. Some of the authors suggest that the difference observed could be explained by greater variance of the measurements during overt speech due to head motion. Another interpretation is that some of these activations result from the inhibition of the articulation process.

Chapter 4

Decoding speech production from electrophysiological activity

As discussed in chapter 1, invasive electrophysiological recording appear as the most suitable method to measure brain activity in the perspective for speech BCIs. Moreover, the majority of recent speech decoding studies relies on ECoG and MEA recordings. In this chapter, the different steps of speech decoding from electrophysiological data are exposed. First the extraction of and selection of relevant neural features is explained. Then, the different decoding strategies are presented. Finally, the performances of speech decoders in existing studies are reviewed.

4.1 Speech neural features

ECoG and MEA recordings allow to record brain activity on different scales. MEA recordings allow to access the firing rates of indiviudal neurons and can also be used to analyze very localized variations of potential in the extracellular (see section 1.2). Recordings using ECoG grids allow to measure the local field potentials (LFPs) on different locations distributed on entire brain regions.

4.1.1 Firing rates

Electrophysiological recordings using MEA allow to record variations of potential in the extracellular medium at the neuronal scale. The neuronal action potentials (or spikes) can be detected using thresholding or pattern matching. A single micro-electrode can record the action potentials of several neurons (usually less than 10). It is possible to attribute the detected spikes to different putative neurons by exploiting the fact that spikes produced by the same neuron tend to have similar waveforms. This operation can be achieved with spike sorting algorithms that require more or less human intervention. Spike sorting is important for the understanding of the behavior of single neurons and provides additional information for neural decoding. However, some studies show that it might not contribute much to BCI performance (Christie et al., 2015).

The successions of events constituted by action potentials are referred to as spike

trains. Since this representation is hardly usable by common decoding algorithms, the spike trains are usually transformed into firing rates by temporal binning or smoothing. In the first case, the time interval containing the spike train is divided into temporal bins in which the number of spikes is counted. In the case of smoothing, a kernel function (for example Gaussian) if convoluted with the spike train (Nawrot et al., 1999). It is also possible to analyze spike trains by analyzing repeated temporal patterns, using techniques such as those developed by (Torre et al., 2016). These methods however have not been used in the context of speech decoding.

4.1.2 Local field potentials

The LFPs measured by ECoG or MEA are usually analyzed in terms of spectral content, typically by computing the variations of power in defined frequency bands, namely: delta (1-4 Hz), theta (4-10 Hz), alpha (7-12 Hz), mu (7-13 Hz), beta (10-30 Hz), gamma (25-80 Hz) and high-gamma (60-200 Hz). These frequency bands have been identified in different brain regions in response to different cognitive tasks or states. Band powers can be extracted using time-frequency transforms like the Fourier transform or the continuous wavelet transform (CWT). It can also be extracted by applying bandpass filtering and computing the envelope of the resulting signal.

Overall, speech production has been associated with an increase in the high gamma (HG) frequency range over motor frontal areas and a more widespread decrease of alpha and beta band powers (Toyoda et al., 2014). HG activity has been shown to be related to cognitive processing and correlated to the neuronal firing rate (Ray and Maunsell, 2011). In several studies, high gamma modulations starting tens of milliseconds before speech onset have been reported in the left premotor, primary motor and sensorimotor cortices as well as Broca's area (Kellis et al., 2010; Pei et al., 2011b,a; Leuthardt et al., 2012; Lotte et al., 2015; Herff et al., 2015). The HG band has been identified as containing relevant information and used in many speech decoding studies, as summarized in Table 2. More specifically, HG activity was found to be related to articulation in the vSMC (Crone et al., 2001; Chartier et al., 2018) and to acoustic features in the auditory cortex (Nourski et al., 2015; Akbari et al., 2019).

Low frequency components of the LFP, like the mu rhythm or under 50 Hz lowpassed activity, have also proven useful to the decoding of speech features (Mugler et al., 2014; Chartier et al., 2018). Low frequency activity has been shown to outperform HG for the decoding of acoustic features from the auditory cortex during speech listening (Nourski et al., 2015; Akbari et al., 2019).

4.2 Dimensionality reduction and features selection

Speech decoding is based on machine learning algorithms whose performance is sensitive to the number of variables in the input data. It has been observed that the predictive power of machine learning models increases with the number of features until an optimal number is reached and that it decreases beyond that point (Hughes, 1968). The first reason (known as curse of dimensionality) is that when the dimension of the data increases,

the distances between the observations grow and the considered dataset becomes sparser. Since statistical analyses become less reliable in sparse data, the number of observations required by most machine learning algorithms to produce reliable predictions grows exponentially with the dimension of the input data. The second reason is that a higher number of input features will often result in models with a higher number of parameters, which are more subject to overfitting. Overfitting refers to the fact that the trained model matches too closely the training set and therefore is not able to generalize to unseen data. To overcome the problems caused by high-dimensional data, speech decoding schemes usually apply dfeature selection and/or dimensionality reduction to the neural data.

Dimensionality reduction refers to the transformation of high-dimensional data into low-dimensional data, with the objective of conserving the relevant properties of the original data. The most common dimensionality reduction technique is the principal components analysis (PCA). PCA projects the original data into a low-dimensional space in such a way that the variance of the projected data is maximized. It was for example used by Kellis et al. (2010) and Wilson et al. (2020) to reduce the dimension of neural time series corresponding to different speech production tasks before classifying them. In other approaches, the transformation is directed by a criterion that is related to the decoding problem. Linear discriminant analysis (LDA) for example computes a linear projection of the original features that maximizes the separability of defined classes in the resulting space. Similarly, the demixed PCA (dPCA) has been developed in the context of neuronal data analysis. It realizes a trade-off between the two types of objectives, i.e. maximizing the captured variance and the separability of the task parameters (Kobak et al., 2016).

The other approach used to reduce the dimensionality of the input data while preserving relevant information is to select a subset of the input variables (features) according to a relevance criterion. In a classification framework, the relevance of a feature can be assessed by comparing the distribution of its values for the different classes. Statistical testing on the distributions' mean values, like Student's t-test or analysis of variance (ANOVA), can be used as a criterion. Before classifying vowels from firing rates, Tankus et al. (2012) for example selected relevant units by comparing their mean firing rates in each class using ANOVA. Mugler et al. (2018) similarly used ANOVA to select the ECoG electrodes and the frequency bands that were to be used in phoneme classification. The distance between the distributions of a feature for the different classes can be measured using more elaborate tools such as the Kullback-Leibler divergence or the mutual information, as in the phoneme classification studies of Herff et al. (2015) and Pei et al. (2011a), respectively. Using criteria such as the ones presented above, it is possible to individually select the most relevant features. In order to avoid selecting redundant features, which increase dimensionality without adding to the model's predictive power, selection procedures considering the selected features collectively have been developed. The minimum redundancy maximum relevance (mRMR) methods aim at selecting a subset of features that fulfill the two criteria. The mRMR method based on information theory developed by Peng et al. (2005) was for example used by Pei et al. (2011a).

4.3 Speech decoding models

There are two main approaches to decode speech from neural activity: the discrete and the continuous decoding approaches. Discrete decoding aim at classifying the recording epochs into a finite set of categories corresponding to the temporal structures of speech, such as phonemes (Brumberg et al., 2011; Pei et al., 2011a; Tankus et al., 2012; Mugler et al., 2014; Song et al., 2014; Ramsey et al., 2018) or words (Kellis et al., 2010; Martin et al., 2016; Makin et al., 2020). The decoded elements can then be used to synthesize speech, for instance using text-to-speech synthesis (see section 2.3.3). On the other hand, continuous decoding approaches directly predict continuous parameters, which can in turn be used by a parametric synthesizer to produce speech (Guenther et al., 2009; Bocquelet et al., 2016c). These continuous parameters can be of two types: acoustic features or articulatory trajectories (see sections 2.3.2 and 2.3.2). In both cases, past studies on speech decoding have usually considered that a sampling rate of 100 to 200 Hz was suitable for the target speech features in order to potentially reconstruct intelligible speech (Bocquelet et al., 2016c; Martin et al., 2014; Anumanchipalli et al., 2019; Akbari et al., 2019). Continuous decoding models are built using machine learning algorithms. These algorithms learn from a training dataset the mathematical operations that transform the neural measurements into the speech features to be decoded. They are then evaluated on another dataset, the test set, to assess their performance. The different models notably differ depending on the amount of data that is required to train them.

4.3.1 Discrete decoding

Discrete decoding uses neural epochs that were recorded during the pronunciation of words, phonemes or other speech segments. The performance of a decoding algorithm is then measured by its ability to classify the neural epochs according to the corresponding segments. An approach to classify neural data epochs is to compute their proximity to other average neural responses: template matching. To classify words using LFP epochs, Kellis et al. (2010) modeled the words as their centroids in the neural feature space and the test epochs were then classified according to their proximity to the centroids. Ramsey et al. (2018) used spatiotemporal matched filters (STMF) to classify words. An STMF is the average neural response accross time on an ECoG grid that is elicited by a specific word. It can be used as a filter to detect new occurrences of the same pattern. The neural responses can also be modeled as probability distribution as in the case of the naive Bayes classifiers used in Pei et al. (2011a) and Herff et al. (2015), where phones are classified based on ECoG data. An extension of traditional template matching approaches is seen in Herff et al. (2019). In this study, the training set was arbitrarily divided into 50000 150 ms-segments of neural data associated with synchronous audio waveforms. It constitutes a database of templates that was used to reconstruct the audio waveform of unseen neural data.

Another way to discriminate the epochs is to build models that project the input features into a space where classes are as separated as possible. This projection can be linear as for Tankus et al. (2012) which used a regularized multivariate linear solver to classify vowels based on firing rates. The linear discriminant analysis (LDA) is commonly

used as a classifier due to its simplicity. Indeed, it has a closed-form solution that is easy to compute and has no parameters to tune. It has successfully been used to classify phonemes based on ECoG data (Mugler et al., 2014). On the other hand, non-linear classifiers can solve a wider range of problems but are more subject to overfitting and often necessitate to subdivide the training set in order to optimize hyperparameters. On a task of phone classification based on firing rates, Brumberg et al. (2011) compared the performance of a linear model, the LDA, with two non-linear models, namely the flexible discriminant analysis (FDA) and the Gaussian kernel support vector machine (SVM). The study found a superiority of the Gaussian kernel SVM. Martin et al. (2016) also used a Gaussian kernel SVM for word-pair classification based on input features derived from high-gamma activity.

4.3.2 Continuous decoding

The most simple tool to decode continuous parameters using neural activity is the linear regression. For example, Martin et al. (2014) fitted a multivariate linear model using gradient descent to predict acoustic speech features based on high-gamma activity. A penalty cost is often added to the loss function of linear models to prevent overfitting. This operation, called regularization, aim at limiting the model's complexity to improve its generalization capabilities. Ridge and lasso regressions are common types of regularized regression models. Conant et al. (2018) for example used a lasso regression model to decode vocal tract movements from high-gamma activity. Partial least squares (PLS) regression algorithms provide another efficient way to avoid overfitting in the decoding of high-dimensional neural data. It projects both the input and the predicted variables in a common low dimensional space and computes a linear regression between the two projections. PLS therefore reduces the dimensionality of the predictor data in a way that is relevant to the regression task. Algorithms of the PLS family have been used in motor BCI studies (Zhao et al., 2013; Eliseyev and Aksenova, 2016) but not yet for speech decoding.

Another approach to decode continuous speech parameters is the use of Kalman filter. Kalman filtering is a recursive algorithm that is designed to optimally estimate a desired variable (the parameter to decode) using the past variations of this variable and synchronous measurements (neural activity). Its strength resides in the fact that it models the relation between the output and the measurements, as well as the dynamics of the output variable. Kalman filtering has shown good performances in cursor control BCI studies (Wu et al., 2006; Kim et al., 2007) and was successfully used to perform online decoding of speech formants from neuronal firing rates (Guenther et al., 2009).

Finally, recent works using ANN have improved decoding results. Chartier et al. (2018) used long short-term memory (LSTM) recurrent neural network to decode estimated articulatory positions from high-gamma activity. The use of such type of network is relevant as it is efficient to deal with time-dependent information (Hochreiter and Schmidhuber, 1997). The chosen architecture however has a quite large number of parameters and is therefore suited to large datasets. In the present example, each subject pronounced 460 sentences. Based on the same approach, Anumanchipalli et al. (2019) used a bidirectional LSTM to decode articulatory trajectories and managed to produce intelligible words after

using an articulatory-to-acoustics speech synthesizer. In the related field of speech decoding during perception, Akbari et al. (2019) managed to decode acoustic parameters and to synthesize intelligible digits from it using not recurrent but feedforward neural networks,.

4.4 Review of speech decoding performances

The present sections discusses the results of 15 speech decoding studies based on electrophysiological data. Details and performances of the different studies are summarized in Table 2.

4.4.1 Discrete decoding

Several studies showed that it is possible to classify limited sets of pronounced words and phonemes based on micro-electrode recordings. Kellis et al. (2010), achieved about 50% accuracy for the classification of 10 words from activity recorded during overt speech using micro-electrodes implanted over the face motor cortex. In a study by Tankus et al. (2012), the firing rates of single cells of 11 patients were used to distinguish 5 vowels with a success rate of 93%. They also noticed that neurons in medial-frontal areas showed a highly specific tuning to individual vowels, whereas STG neurons had nonspecific, sinusoidally modulated tuning (analogous to motor cortical directional tuning). In the study of Wilson et al. (2020), phonemes could surprisingly be decoded using a Utah array inserted in the "hand knob" area of the dorsal motor cortex. High frequency power (125-5000 Hz) was found more effective than firing rates for decoding.

Significant classification accuracy was also obtained using ECoG recordings. In a study by Mugler et al. (2014), on average 20% of accuracy was achieved at decoding 31 different English phones (thus with a chance level of 7.4%) from ECoG data recorded from 4 subjects overtly reading isolated monosyllabic words. In Ramsey et al. (2018), the classification between 4 phones reached 75% on average across 5 subjects with most informative locations being along he central sulcus.

Words and phonemes have also been decoded from experiments involving covert speech and persons suffering from locked-in syndrome. In Pei et al. (2011a) phones were decoded not only from overt but also from covert speech. With a chance level of 25%, vowels and consonants classification yielded respectively 41% for overt speech and about 37% for covert speech. Using micro-electrodes in a patient suffering from locked-in syndrome and thus who could not produce any movement, 38 different American English phones could be decoded with about 20% accuracy (Brumberg et al., 2011). In the studies of Brumberg et al. (2011) and Ramsey et al. (2018), as opposed to the one by Pei et al. (2011a), consonants were found to be more reliably decoded than vowels. Ikeda et al. (2014) were able to classify 3 vowels over chance based on the neural activity of participants performing articulated covert speech (also called mimed speech). They did so using each time the power in a single frequency band (among theta, alpha, beta, high-gamma) at a single electrode (mostly in the left temporal and frontal gyri). In Martin et al. (2016), word-pair classification reached 88% for imagined speech with a mean of 58% across five subjects. For overt speech and speech listening, 86 and 89% were respectively obtained.

Study	Device	Area	Activity	Representation	Performance
Guenther et al., 2009	MEA	vSMC	smoothed FR	(C) formants	Correlation: r $\simeq$ 0.5 (covert)
Kellis et al., 2010	μ-ECoG	vSMC Wernicke	0-500 Hz PSpg	(D) words	2-class accuracy: 85%
Pei et al., 2011a	ECoG	large	0-200 Hz PSpg	(D) phonemes	4-class accuracy: $\sim$30% (overt) $\sim$40% (covert)
Brumberg et al., 2011	MEA	vSMC	binned FR	(D) phonemes	24-class accuracy: 21%
Tankus et al., 2012	MEA	vSMC ACC	binned FR	(D) vowels	5-class accuracy: 93% (covert)
Ikeda et al., 2014	ECoG	large	HG power (70-110 Hz) beta power (14-30 Hz) alpha power (8-13 Hz) theta power (4-7 Hz)	(D) vowels	3-class accuracy 42-56% (covert)
Mugler et al., 2014	ECoG	vSMC	HG power (65-250 Hz) mu power (7-13 Hz)	(D) phonemes	31-class accuracy: 20%
Martin et al., 2014	ECoG	vSMC STG Wernicke	HG power (70-150 Hz)	(C) audio PSpg	Correlation: r = 0.41 (overt) r > chance (covert)
Herff et al., 2015	ECoG	large	HG power (70-170 Hz)	(D) phonemes	21-class accuracy: 10-50%
Martin et al., 2016	ECoG	large	HG power (70-150 Hz	(D) words	2-class accuracy: 86% (overt) 58% (covert)
Chartier et al., 2018	ECoG	vSMC	HG power (70-150 Hz)	(D) articulation	Correlation: r = 0.43
Ramsey et al., 2018	ECoG	vSMC	HG power (65-125 Hz)	(D) phonemes	4-class accuracy: 30%
Herff et al., 2019	ECoG	vSMC IFG	HG power (70-170 Hz)	(D+S) speech units	4-word accuracy: 66%
Anumanchipalli et al., 2019	ECoG	vSMC STG IFG	HG power (70-200 Hz) 0-30 Hz activity	(C+S) articulation	transcription: 53% WER
Wilson et al., 2020	MEA	dMC	125-5000 Hz power	(D) phonemes	39-class accuracy: 30%

Table 2. Summary of speech decoding studies based on electrophysiological data. (Area) The indicated brain areas are situated on the left hemisphere for a large majority of the participants of the featured studies. The area is defined as "wide" when the ECoG coverage of the study encompasses a large part of the left hemisphere including the IFG, the vSMC and the STG. ACC: anterior cingulate cortex; dMC, dorsal motor cortex; HeG: Heschel's gyrus. **(Activity)** The neural features that did not significantly contribute to decoding were not included. FR: firing rates; HG: High-gamma. PSpg: power spectrogram. **(Representation)** This column indicates the decoded speech representations. C and D are used to mark continuous and discrete decodings. S indicates that speech was synthesized using the decoded representation. **(Performance)** This column summarizes the decoding performance, mostly in terms of correlation (Pearson's correlation coefficient) and classification accuracy. For Martin et al. (2014), the covert correlation was superior to chance level but cannot be compared to other correlation coefficients due to the use of dynamic time warping. For the studies featuring speech synthesis, subjective evaluations are provided. For Anumanchipalli et al. (2019), the word error rate (WER) of transcribed sentences (50-word vocabulary) is indicated.

In a study by Herff et al. (2015) based on ECoG, prior linguistic knowledge was added to phone decoding by limiting the vocabulary to 10 words and building a statistical model giving the occurrence of a word given the preceding one. While raw phone recognition accuracy ranged from about 10% to 50% (depending on subject and session) for a total of 20 phones in an overt reading task, the use of a statistical model allowed to reach a word recognition accuracy of about 75%. Finally, Herff et al. (2020) were able to synthesize speech by decoding 150 ms speech units from ECoG. Within 4-word groups, listeners were able to identify the word originally pronounced by the subject with 66% accuracy.

4.4.2 Continuous decoding

In the study of Guenther et al. (2009), the activity recorded from intracortical microelectrodes localized in the ventral sensorimotor was used by a locked-in patient to control in real-time a speech synthesizer in order to produce vowels and transitions between them. Spike sorting was applied on the recordings and firing rates of the putative cells were used to estimate the two first formants (see section 2.1.3) of the speech signal. This study constitutes the only reported case of closed-loop speech BCI.

In Martin et al. (2014) a spectro-temporal representation of sounds (obtained using wavelet transform on the audio signal with 32 logarithmically-spaced frequency bins between 180Hz and 7kHz) was directly inferred from ECoG recordings, both during overt and covert speech. Although this approach could not produce intelligible speech, the overall time-frequency structure of the speech spectrograms could be estimated above chance level.

In Chartier et al. (2018), ECoG recordings in the vSMC were analyzed in regard with articulatory trajectories estimated by acoustic-to-articulatory inversion (see section 2.1.2). Interestingly, the study found that coordinated articulatory movements, like "lip closing", are encoded in single electrode activity, rather than movements of single articulators (like the upper lip position). The articulatory trajectories decoded from the ECoG high-gamma were significantly correlated with the trajectories inferred from the audio data with an average correlation coefficient of 0.43. Based on the same approach, Anumanchipalli et al. (2019) could decode articulatory trajectories with a precision that was high enough to produce intelligible speech synthesis for one subject. The decoded sentences were transcribed by human listeners and a 53% word error rate (WER) was obtained when considering a 50-word vocabulary. Their study showed an advantage of decoding articulatory trajectories before inferring acoustics over the direct prediction of acoustics. They found that the vSMC contributed more to the decoding than the IFG and the STG. They also applied the model trained on overt speech to mimed speech data and obtained over chance results, confirming that the activity supporting the decoding does not only rely on the auditory feedback of the participants' vocalization.

Chapter 5

Acoustic contamination of electrophysiological recordings

5.1 Introduction

Motion artifacts are classically seen in electrophysiological signals. In particular, mechanical vibrations may create variations of biopotential measurements (Luna-Lozano and Pallas-Areny, 2010). Such undesired signals may have different origins, including the bending of the electrode wires and the electrochemical changes at the electrode-electrolyte interface induced by small displacements of the electrodes (Salatino et al., 2017; Nicolai et al., 2018). Similarly, we found that the impact of sound waves on the acquisition system can cause a contamination of the electrophysiological recordings by the acoustic signals. This phenomenon, which had not yet been reported to our knowledge, has important implications for the studies based on electrophysiological recordings.

As summarized in Table 2, most of the recent studies decoding speech production from neural activity use field potential signals in the high-gamma frequency range, which typically covers frequencies from 70 to 200 Hz. This neural features has also been used for the decoding of speech stimuli (Pasley et al., 2012; Akbari et al., 2019). A noticeable feature of acoustic speech signals is the fundamental frequency f_0 of the human voice, which corresponds to the vibrational source of speech produced by the vocal folds in the larynx and further modulated by the vocal tract to produce the variety of speech sounds (see section 2.1.2 for more details). The fundamental frequency depends on the size of the vocal folds and typically falls around 125 Hz for males and 215 Hz for women (Small, 2012). The high-gamma frequency band and the range of the fundamental speech frequency thus generally overlap. At frequencies above f_0, the acoustic content of speech is further characterized by the harmonics of the fundamental frequency, which typically overlap frequencies of unit and multi-unit neural activity.

Here, we analyzed human ECoG and intracortical recordings during speech production and perception as well as a rat μ-ECoG recording during sound perception. We found that electrophysiological signals may often be contaminated by spectrotemporal features of the sound produced by the participant's voice or played by the loudspeaker. This contamination seems to result from a microphonic effect at the level of the cables and

connectors along the recording chain, affecting the range of high-gamma frequencies and above. These findings suggest acoustic contamination could bias decoding performances in some of the past and future speech decoding studies and therefore negatively impact the development of speech BCIs. To prevent this, care should be taken to exclude such artifacts when investigating cortical signals underlying speech production and perception. The results exposed in this chapter have been published in Roussel et al. (2020).

5.2 Neural datasets

The datasets described in this section are divided between those that were obtained from the Brainspeak research project, additional datasets that were provided by collaborators and a rat recording. The datasets obtained from the Brainspeak project feature recordings from participants P2, P3 and P5. The recordings of participants P3 and P5 have also been used in chapters 6 and 7. The additional datasets and the rat recording have solely been used in the study of acoustic contamination in this chapter.

The recordings of P2 and P3 were realized by Blaise Yvert's team prior to the beginning of my thesis. I was actively involved in all the recordings of participant P5. The rat recording has been performed as part of another study by Marie Palma, PhD student in the same research team.

5.2.1 Brainspeak datasets

5.2.1.1 Participants

The results of the present thesis are primarily based on electrophysiological recordings obtained in 3 patients at the Grenoble-Alpes University Hospital (*Centre Hospitalier Universitaire Grenoble-Alpes*, CHUGA) as part of the Brainspeak clinical trial (NCT02783391) approved by the French regulatory agency ANSM (DMDPT-TECH/MM/2015-A00108-41) and the local ethical committee (CPP-15-CHUG-12). The neural recordings of P2 and P3, a 42-year-old male and a 29-year-old male respectively, was performed during their awake surgery for tumor resection. P5, a 38-year-old female, was implanted for 7 days as part of a presurgical evaluation of her intractable epilepsy. These 3 participants gave their informed consent to participate in the study. All surgeries were performed by Pr Stéphan Chabardès.

5.2.1.2 Electrophysiological recordings

Brain activity from participants P2 was recorded in the operating room during awake surgery prior to tumor resection. A 256-electrode array (PMT Corp., USA) was positioned after opening the skull and the dura matter over the left sensorimotor cortex and the tumor (figure 11). Ground and reference electrodes were integrated on the back side of the array and maintained wet using compresses soaked with saline. The 16 electrodes' pigtails were connected to eight 32-channels Cabrio Connectors (Blackrock Microsystems, USA) connected by shielded cables to two front-end amplifiers (FEA, Blackrock Microsystems, USA). In the FEA, the signals were amplified and digitized at 30 kHz. The digitized signals

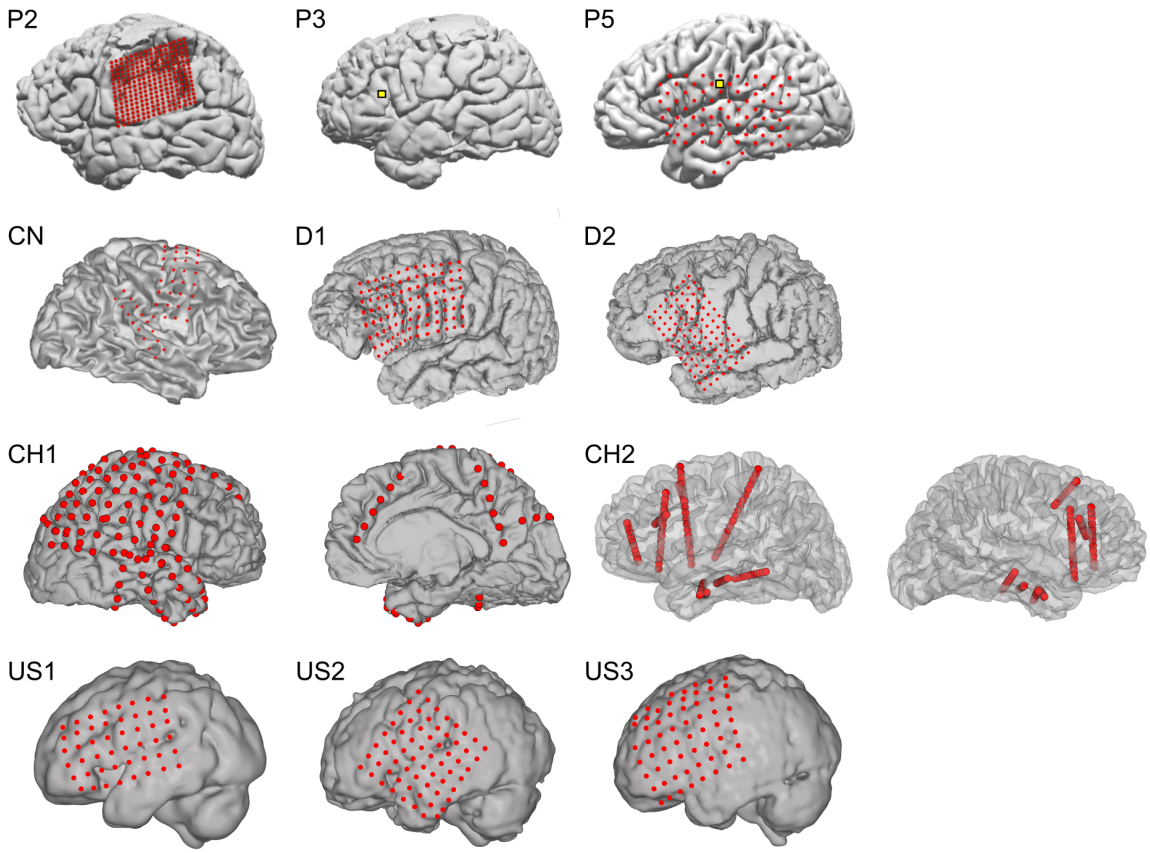


Figure 11. Electrode placement for the different human participants. Red: ECoG electrodes (all but P3). Yellow: intracortical Utah array (P3 and P5).

were then transmitted by an optic fiber to two synchronized Neural Signal Processors (NSP, Blackrock Microsystems, USA) interfaced with a computer and filtered between 0.3 and 2500 Hz.

Brain activity from participants P3 was recorded in the operating room during awake surgery prior to tumor resection. A 96-channel intracortical Utah microelectrode array (UEA, Blackrock Microsystems, USA) with 1.5-mm long electrodes was inserted in the pars triangularis of Broca's area (figure 11), at a location that was subsequently resected to access the tumor for its removal. The pedestal serving as ground was screwed to the skull. Two wires with deinsulated tips were inserted below the dura, and one was used as reference. The electrodes were connected via a Patient Cable (Blackrock Microsystems, USA) to a FEA where signals were digitized at 30 kHz, and further transmitted through an optic fiber to a NSP. The signals were filtered between 0.3 and 7500 Hz.

Brain activity from participant P5 was recorded in subchronic condition at the hospital. Participant P5 was implanted with a 72-electrode ECoG array (PMT Corp., USA) covering a large portion of her left hemisphere as well as a 4-electrode strip (PMT Corp., USA) over the left ventral temporal lobe and a 96-electrode UEA inserted in the left ventral sensorimotor cortex (figure 11). An electrode of the strip was used as the reference and

another as the ground. The transcutaneous pigtails of the ECoG grids were connected to PMT pigtail adaptors and then to two headboxes (64-Channel Splitter Box, Blackrock Microsystems, USA) through individual touch-proof connectors. The headboxes were connected to a FEA linked to a NSP. The transcutaneous pedestal of the UEA was screwed to the skull and connected to a Cereplex-E headstage (Blackrock Microsystems, USA) ensuring signal amplification and digitization before transmission to a second NSP through a digital hub. For these intracortical recordings, the reference was a wire deinsulated at its tip and inserted below the dura, and the ground was the pedestal. Data from both electrode arrays was sampled at 30 kHz recorded on the two synchronized NSPs. The signals were filtered between 0.3 and 7500 Hz.

5.2.1.3 Audio recordings

In the case of speech production tasks, the participant's speech was recorded along with his or her neural data. For participants P2, P3 and P5, a microphone (SHURE Beta 58 A) was positioned at about 10-20 cm from the mouth. The signal was amplified using an audio interface (Roland OCTA-CAPTURE) and digitized by one of the NSPs, at the same rate and synchronously with the neural data (as in Figure 12).

5.2.1.4 Tasks and stimuli

Participants P2, P3 and P5 were asked to read aloud short French sentences. All three participants were native French speakers. In the considered sessions, P2, P3 and P5 pronounced 118, 86 and 97 sentences, respectively.

Participant P5 also took part in a closed-loop protocol involving speech perception, where she was exposed to the sound of computer-generated phonemes delivered by a loudspeaker positioned about 50 cm on her left. This protocol is further described in Chapter 7. This second dataset involving P5 also contains speech production segments as she was interacting with the experimenters. The first (speech production) and the second (speech perception and production) datasets are labeled as sessions a and b, respectively.

5.2.1.5 Closed-loop recordings

Participant P5 also took part in a closed-loop protocol. She was asked to imagine producing sustained sounds with the aim of controlling a speech synthesizer. Two of the closed-loop sessions have been used for the study of acoustic contamination. The participant was exposed to the sound of computer-generated vowels delivered by a loudspeaker positioned about 50 cm on her left. These recordings also contains speech production segments as the participant was interacting with the experimenters. The first (speech production) and the second (speech perception and production) datasets are labeled as sessions a and b, respectively.

5.2.2 Additional human datasets

5.2.2.1 Participants

As part of the study on acoustic contamination, we further included datasets obtained by 4 other centers in China, Germany, Switzerland and the USA. First, a 22-year-old male participant (CN) suffering from intractable epilepsy requiring surgical treatment was recorded at the Second Affiliated Hospital of Zhejiang University. These procedures were followed from the guide and approved by the Second Affiliated Hospital of Zhejiang University, China. Participant CN gave written informed consent after detailed explanation of the potential risks of the research experiment. Second, two additional datasets were acquired at Frankfurt University. A 33-year-old bilingual Russian and English-speaking male patient suffering from a left frontal anaplastic astrocytoma (D1) and a 36-year-old native German speaking female suffering from an anaplastic glioma (D2) were recorded. Participants D1 and D2 gave written informed consent after detailed explanation of the potential risks of the research experiment, which were approved by the ethics committee of the medical faculty of Goethe University (GZ 310/11). Thirdly, two patients with drug-resistant epilepsy were recorded extra-operatively at Geneva University Hospitals: a 49-year-old woman (CH1) and a 50-year-old man (CH2). Both gave written consent to participate in speech production and processing experiments, which were approved by the local ethics committee (Commission cantonale d'éthique de la recherche, Geneva, Switzerland). Finally, datasets were recorded from three patients, a woman aged 18 (US1), a man aged 42 (US2), and a man aged 21 (US3), at the University of Washington (Seattle, WA, USA). All 3 patients participated in a purely voluntary manner, after providing informed written consent, under experimental protocols approved by the Institutional Review Board of the University of Washington (#12193). Patient data was anonymized according to IRB protocol, in accordance with HIPAA mandate (Miller, 2019).

5.2.2.2 Electrophysiological recordings

Brain activity from participants D1 and D2 was recorded in the operating room during awake surgery prior to tumor resection. Participant D1 received left perisylvian electrocorticography using an 8x12 electrode grid and two 2x8 electrode grids with 5mm spacing (Ad-Tech Medical, USA; see figure 11). The ground electrode was the subgaleal needle electrode P4 and data were referenced against a frontocentral subgaleal needle electrode Fz and sampled at 5 kHz (four amplifiers connected to two headboxes; BrainAmp MR plus amplifier, BrainProducts, Germany). Participant D2 received left perisylvian electrocorticography using an 8x12 grid electrode (Ad-Tech Medical, USA). The ground electrode was the subgaleal needle electrode C4 and data were referenced against a frontocentral subgaleal needle electrode Fz and sampled at 2048 Hz (two amplifiers 64 channels each, one headbox per amplifier, Micromed, Italy).

Brain activity from participants CN, CH1, CH2, US1, US2, and US3 was recorded in subchronic condition at the hospital. Participant CN was implanted with a 32-electrode clinical subdural ECoG grid (HuakeHesheng, China) in his right sensorimotor cortex for clinical monitoring and localization of his seizure foci (Figure 1). The clinical electrodes were platinum electrodes with a diameter of 4 mm (2.3 mm exposed) spaced every 10 mm,

implanted for seven days. The configuration and location of the electrodes, as well as the duration of the implantation, were determined by clinical requirements. An electrode of the ECoG grid was used as the reference and another as the ground. The transcutaneous pigtails of the ECoG grids were connected to HuakeHesheng adaptors. The adaptors were connected to a headbox linked to a NSP via an FEA, like for P5. The signals were sampled at 30 kHz and filtered between 0.3 and 7500 Hz. Participant CH1 was implanted with grids and strips of subdural electrodes over the right cerebral hemisphere (124 recording sites; Ad-Tech Medical, USA). Participant CH2 was implanted with depth electrodes through stereotaxic surgery in both cerebral hemispheres (213 recording sites; Dixi Medical, France). EEG signals for CH1 and CH2 were referenced to a subdermal wire electrode inserted at the vertex, digitized at 2048 Hz and recorded with a BrainQuick LTM system (Micromed, Italy). The Seattle participants all had subdural platinum electrode arrays (Ad-Tech Medical, USA) implanted over the left hemisphere in 6x8 (US1) and 8x8 (US2 and US3) rectangular arrays. These electrodes had 4mm diameter (2.3mm exposed), 1 cm inter-electrode distance, and were embedded in silastic. These data were recorded with Synamps2 amplifiers (Compumedics Neuroscan, USA) in parallel with clinical recording, sampled at 2kHz (US1) or 1kHz (US2 and US3). The data were exported from the amplifiers to the BCI2000 software environment on a separate laptop, using a TCP/IP protocol (Schalk et al., 2004).

5.2.2.3 Audio recordings

For participant CN, the played sentences were realigned with the neural signals using triggers indicating the start of the stimuli. D1's and D2's voices were recorded using a custom-made microphone (BrainProducts, Germany) connected to two bipolar EMG channels of the headboxes. For participants CH1 and CH2, the patient's produced speech or the sound delivered by the stimulus presentation portable computer positioned in front of the patient was captured with a battery-powered microphone (TCM160, AV-Leader Corp, Taiwan) and fed to an available analog input in the EEG amplifier through a custom-made jack-to-touchproof cable. For US1, US2 and US3, speech was recorded using a Logitech USB microphone (Logitech, Lausanne, Switzerland) placed approximately 20cm from the patients' mouth, input to a USB port on a laptop separate from the amplifiers, where sound sample indices were logged into the BCI2000 programming environment (Schalk et al., 2004).

5.2.2.4 Tasks and stimuli

Participant CN was asked to listen and repeat aloud individual sentences of an ancient Chinese poem. Each block consisted of 4 sentences and each sentence lasted between 2 and 5 seconds. There were six blocks in total, 3 from the morning and 3 from the afternoon of the same day. The two sessions are labeled as a.m. and p.m. Participants D1 and D2 performed a sentence repetition task which consisted of an auditory presentation of pre-recorded sentences and their repetition following a visual go signal. The task and stimuli are described in (Gehrig et al., 2019). Participant CH1 performed a speech production task, where she had to repeat a written word after a 2-second delay, and a speech

processing task, where she heard fragments of a presidential discourse. Participant CH2 performed a speech perception task, where he heard fragments of movie soundtracks that contained speech. Participants US1, US2 and US3 performed a simple verb-generation task, where nouns (approximately 2.5 cm high, and 8–12 cm wide) were presented on a screen approximately 1 m from the patient, at the bedside. The patient's task was to speak a verb that was connected to the noun: for example, if the cue read "ball", the patient might say "kick", or if the cue read "bee", the patient might say "fly". In between each 1.6-second cue was a blank-screen 1.6-second inter-stimulus interval (task and stimuli described in further detail in Miller et al. (2011)).

5.2.3 Rat recording

An electrophysiological recording of a rat during sound perception has also been used for the study of acoustic contamination. The surgery and recording have been performed by Marie Palma, PhD student in the same research team. It was obtained over the left auditory cortex of a ketamine (90 mg/kg)-xylazine (2 mg/kg) anesthetized 600-gram adult Sprague Dawley rat using a 64-electrode µ-ECoG array (E64-500-20-60-H64; NeuroNexus Inc, USA). These data were obtained in compliance with European (2010-63-EU) and French (decree 2013-118 of rural code articles R214-87 to R214-126) regulations on animal experiments, following the approval of the local Grenoble ethical committee ComEth C2EA-12 and the ministry authorization 04815-02. A bone screw was used for the ground and a stainless-steel wire inserted below the skin ahead of Bregma was used for the reference. Signals were acquired using the RHD2000 acquisition system and two 32-channel RHD2132 headstages (Intan Technologies, USA). To avoid any possible crosstalk inside the Intan acquisition system, the sounds delivered to the rat were recorded on an independent CED Micro1401 (Cambridge Electronic Design, UK). Both acquisition devices were interfaced and synchronized by the Spike2 software with the IntanTalker module (CED programs) and signals were digitized at 33.3 kHz. The time jitter between sound and neural signals was checked to be below 2 ms. Pure tones (3-ms rise, 167-ms plateau and 30-ms fall times) with frequencies ranging from 0.5 to 16kHz were presented with pseudo-random inter-stimulus intervals of 1.8-2.2 seconds. Sounds were delivered at about 80-90 dB SPL in open field configuration using a MF1-S speaker (Tucker Davis Technology Inc, USA). The three lowest tone frequencies that were further considered in the present study are 0.5, 1 and 2.5 kHz.

5.3 Methods

5.3.1 Pre-processing

5.3.1.1 Audio preprocessing

To center audio signals, a moving average was computed over 1-second windows and subtracted.

5.3.1.2 Segments selection

Whenever the audio recordings contained data from more than one source (stimuli, participant's voice, experimenter's voice), a single source was studied at a time by manually annotating and excluding the segments featuring the other source(s).

5.3.1.3 Channel selection

All electrophysiological recordings were visually inspected. Channels showing signals having unexpectedly high or low amplitude throughout the whole recording were discarded. For participant P2, 112 electrodes were removed due to several loose connections at the level of the Cabrio Connectors. For participant P5, 1 noisy ECoG electrode was excluded.

5.3.1.4 Robust estimation of standard deviation

In the case of normal data contaminated by samples with high absolute values, potential outliers can be detected using a threshold based on standard deviation. For example, samples exceeding 5 times the standard deviation of the whole time series can be consiered as outliers as a precaution.

However, the presence of outliers can significantly bias the estimation of the standard deviation of the rest of the samples. To overcome this problem, the standard deviation can be estimated based on the median absolute deviation (MAD). In the hypothesis of a normally distributed random variable, the standard deviation can be estimated from the MAD using the following formula (Hampel, 1974):

$$\sigma = \frac{MAD}{\sqrt{2}erf^{-1}(1/2)} \approx 1.4826 \cdot MAD$$

with the MAD defined as:

$$MAD(X) = median(|X_i - median(X)|)$$

For approximately normal data, this estimate is robust to the presence of outliers in the data (Leys et al., 2013). It is therefore useful in order to fix thresholds for the detection of high-amplitude artifacts.

5.3.1.5 Artifacts exclusion

High-amplitude artifacts can significantly affect statistical analyses on electrophysiological recordings, in particular correlations. As a precaution, time segments containing abnormally high values were therefore excluded from the analyses. First, the electrophysiological signals were detrended by subtracting their 500-ms moving average. Samples with abnormally high values were then detected by applying a threshold on the absolute values of the recorded potential. This threshold was set individually for each channel to 5 times the standard deviation, estimated using MAD (see preceding section). A time sample was considered as part of a recording artifact when at least 10% of the channels displayed

abnormally high values. A window of 500 ms was excluded for all channels before and after each these samples.

5.3.1.6 Common average reference

Electrophysiological signals contain noise that is common to all channels and noise that is specific to individual channels. In order to remove the common noise in multi-electrode recordings (e.g. power line noise), re-referencing is often used, in particular common average reference (CAR). Common average reference consists in subtracting to each channel the average neural signal, computed on the signals of all selected channels.

Common average reference was applied on the recording of participants P3 to lower the influence of the intrinsic spatial correlation of LFPs stemming from the close spacing of the electrodes of the Utah array, and of participant D1 for whom the reference electrode was itself very noisy, which affected all channels in a similar way. CAR was also applied on participant P2 recording to analyze its effect as shown in Figure 12.

5.3.2 Spectrogram computation

In the present study, a spectrogram refers to the time-varying power spectral density (PSD) computed over a recording channel. For all analyses, spectrograms of neural and audio data were computed using short-time Fourier transform with 200-ms time windows (after Hamming windowing). The window overlap was chosen to obtain spectrograms sampled at 50 Hz. The mean sound PSD (or mean power spectrum) of a recording was computed by averaging the audio spectrogram over all time samples (after selection like described in section 5.3.1.2). For display purposes, the spectrograms in Figures 12, 13 and 20 were computed with higher frequency and time resolutions. These spectrograms were also z-scored within each frequency bin using artifact-free data segments containing the displayed extracts.

5.3.3 Audio-neural correlations

For all recordings, the correlations between the neural and the audio spectrograms were computed. For each channel, the sample Pearson correlation coefficient r between the power amplitudes across time of the channel and audio signals was computed for all possible pairs of frequency bins, resulting in an audio-neural correlation matrix. Correlations corresponding to the same frequency bin between the two signals (i.e., the diagonal of the correlation matrix) are further termed audio-neural correlations. For each value of r, a p-value was computed using Student's t-test to test the null hypothesis that $r = 0$. For audio-neural correlations, the statistical significance of each correlation coefficient was determined with respect to a Bonferroni adjusted significance level $\alpha = 0.05/N$ where N was the number of frequency bins times the number of channels in the recording.

5.3.4 Objective assessment of contamination

To determine whether a dataset is contaminated, we developed a specific statistical approach. First, a contamination matrix was built by computing the maximum of the correlation matrices across all electrodes. Each maximum was thus computed separately for each element of the matrix (i.e. each pair of frequency bins). Then, we evaluated the values on the diagonal of the contamination matrix in relation to the rest of the matrix. The mean value on the diagonal was computed to obtain a contamination index. To evaluate the statistical significance of this original index, a distribution of surrogate index was then built by computing the mean diagonal 10000 times on as many shuffled versions of the contamination matrix. Each shuffled matrix was built by randomly shuffling either the lines or the columns. Shuffling only one dimension at a time favors the preservation of the values on the diagonal in case of horizontal or vertical patterns in the original matrix. The original contamination index was finally compared to the distribution of the surrogate ones. The proportion P of surrogate indices that were superior to the original index was considered as the risk taken when rejecting the null hypothesis that no contamination exists (in other words P was considered as the probability of being wrong when considering that a contamination exists).

5.3.5 Neural decoding

The analyses in this section have been carried out by Gaël Le Godais, PhD student in the same research team. ECoG data from participants P2 and P5 were used to predict acoustic mel-cepstral coefficients of overt speech produced by the participants. Both participants were visually presented with a series of short sentences or vowel sequences written on a screen positioned about 50-100 cm in front of them, and asked to repeat them overtly. The number of sentences was 118 for participant P2 and 150 for participant P5, corresponding to an overall duration of 230 and 329 seconds of speech, respectively. The participants' speech audio signals were decomposed into 25 mel-cepstral coefficients (MELs; see section 2.1.3.2) using the *mcep* function of the SPTK toolkit (http://sp-tk.sourceforge.net/). Spectrograms of the ECoG data were computed as described in paragraph 5.3.1.2 but at a rate of 100 Hz. Neural features were the spectrogram amplitudes in 10-Hz bands (i.e. 0-10, 10-20, ... 190-200 Hz) and the band-pass filtered time domain LFP signal (between 0.5 and 5 Hz). Two sets of neural features were considered, a first one where only features below 90 Hz were used and a second one where all features up to 200 Hz were used. A feature selection process was applied to keep only the features that were significantly modulated during speech production with respect to silence intervals, as assessed by Welch's t-test with a Bonferroni adjusted significance level $\alpha = 0.05/N$ where N is the number of electrodes times the number of candidate features. The resulting number of selected features was 3147 out of 5376 for P2 and 1115 out of 1512 for P5. These selected features were normalized and decomposed using PCA (both transformations were based on training sets). The first 50 (for the 0-90 Hz feature set) or 100 (for the 0-200 Hz feature set) components were used as the final set of features. A linear model was then used to map these neural features onto the mel-cepstral trajectories using 10-fold cross-validation. Each mel-cepstral sample was decoded using a 200 ms window of neural activity centered

on the time of this sample. Chance decoding level was assessed by repeating the whole procedure after shuffling and time-reversing the mel-cepstral trajectories of the different sentences (truncation was applied to match sentence durations).

5.4 Results

5.4.1 Observation of acoustic contamination in neural recordings

5.4.1.1 Correlation between ECoG and sound signals during speech production

We observed strong correlations between ECoG and sound spectrograms in participant P2 during speech production. Participant P2's brain activity was recorded with an ECoG grid while he was reading sentences aloud. Simultaneously, a microphone was used to capture the sound of his voice (see Figure 12a). Figure 12b shows a portion of the z-scored spectrograms of the sound signal (top) and of an electrode of the ECoG grid (bottom). In this example, the ECoG signal shows a very similar spectrotemporal structure as that of the sound. The time-frequency patterns observed are consistent with human speech and are unlikely to be brain activity.

We quantitatively assessed this phenomenon by computing the correlation between the power of the signal within each frequency bin of each electrode signal with that of the sound signal. As shown in Figure 12c and in the top of Figure 12d, correlations up to 0.6 could be observed depending on the electrode. Up to 370 Hz, the strongest correlations were observed at frequencies most present in the sound signal, and in particular between 115 and 145 Hz, which corresponded to the range of the fundamental frequency of the subject's voice. Above 370 Hz, correlations were low even at frequencies for which the power of the speech signal remained high. As shown in Figure 12d, the correlations between sound and ECoG spectrograms were still present and even exacerbated after common average re-referencing of the ECoG signals.

5.4.1.2 Correlation between intracortical and sound signals during speech production

In P3 recording, we further observed statistically significant correlations between the spectrograms of intracortical signals recorded using a Utah array and that of the produced speech signal. Figure 13a shows a portion of the z-scored spectrograms of the subject's voice (top) and of one electrode of the array (bottom). The spectrogram of the selected micro-electrode clearly shows spatio-temporal features also observed in the sound spectrogram (between 200 Hz and 400 Hz). Statistically significant correlation coefficients up to 0.7 were observed, with peaks falling in the range of frequencies where the sound signal showed high power (Figure 13b). Noticeably, correlations between intracortical and sound signals during speech production were much weaker in participant P5 (Figure 14b).

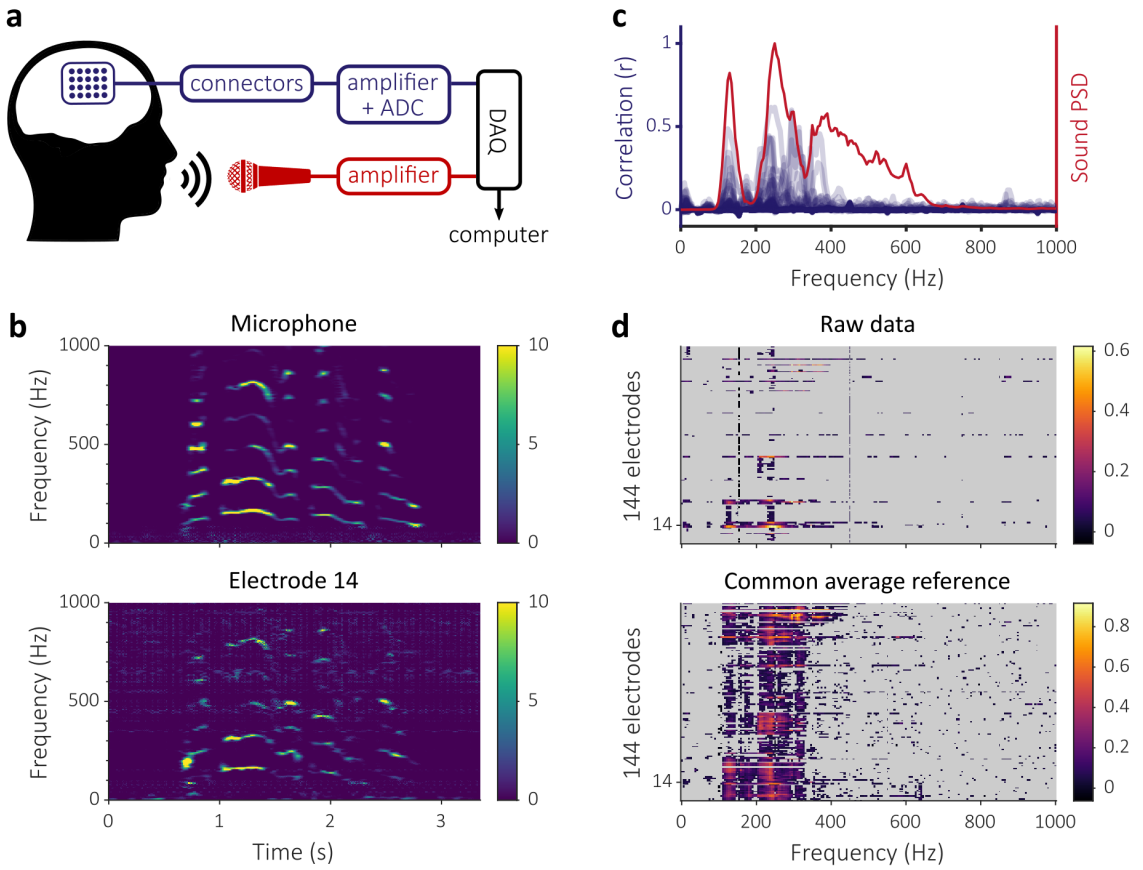


Figure 12. Correlation between voice and ECoG signals during speech production in participant P2. (a) Schematic representation of the recording setup, including neural (blue) and audio (red) data streams. The analog-to-digital conversion (ADC) of the audio signal is done in the data acquisition system (DAQ) whereas it in the FEA for neural signals (see section 5.2 for more details). **(b)** The upper and lower graphs show the z-scored spectrograms of the microphone and of electrode 14, respectively. The succession of stable striped patterns interleaved by transient states is typical of human speech formants. **(c)** Each blue curve represents, at all frequency bins, the value of the correlation coefficients between the spectrogram of one electrode signal and the spectrogram of the audio signal. The red curve represents the mean PSD of the audio signal (a.u.). **(d)** Heat maps representing the correlation coefficients between audio and neural data across electrodes and frequency bins. Correlation coefficients not statistically significant are displayed in grey. The upper and lower graphs show the results when using raw neural data and neural data after common average reference, respectively.

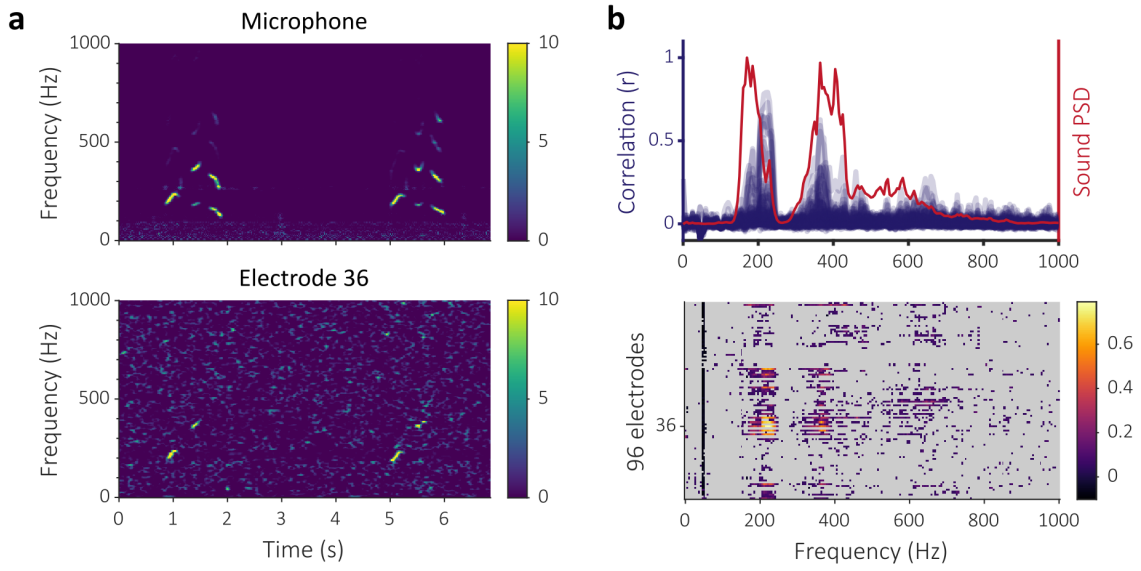


Figure 13. Correlations between voice and intracortical signals during speech production in participant P3. (a) The upper and lower graphs show the z-scored spectrograms of the microphone and electrode 36, respectively. **(b)** On the upper graph, each blue curve represents, at all frequency bins, the value of the correlation coefficients between the spectrogram of one electrode signal and the spectrogram of the audio signal. The red curve represents the mean PSD of the audio signal (a.u.). The lower panel represents a heat map of the correlation coefficient between audio and neural data for all electrodes and frequency bins. Correlation coefficients not statistically significant are displayed in grey.

5.4.1.3 Correlation between electrode and sound signals during sound perception

Statistically significant correlations between electrode and sound signals were not only present during speech production as reported above, but also during sound perception. This phenomenon was observed in human and animal recordings using different recording instrumentations. Participant P5 participated during session b in a paradigm where artificially synthesized speech sounds were presented to her through a loudspeaker positioned on her left. Brain activity was recorded from both ECoG electrodes and intracortical microelectrodes. The sound produced by the loudspeaker was also recorded simultaneously. During this session, the segments when the participant was speaking to the experimenter were also separately studied. Performing the same analysis as in the previous section, we found that ECoG signals during speech production segments were not as highly correlated with the participant's voice (Figure 14a and 14b). The highest correlations were found in the ECoG recording, within the frequency range of the voice's first harmonic (Figure 14a). The neural recordings showed strong correlations with the perceived sound signal, with peaks up to 0.9 (Figure 14c). As observed in the recordings from P2 and P3, frequencies showing strong correlations were mostly found in the bands that concentrate most of

the sound power. By comparison, the spectrograms of intracortical signals were poorly correlated with that of the sound (Figure 14d).

Second, in order to verify that the correlations were not due to our clinical recording system in particular, we performed the same type of analysis on data obtained from an experiment in a rat. The left auditory cortex was recorded using a commercial µ-ECoG grid connected to an Intan neural recording system (Figure 15a). In this case, pure tones were delivered in an open field configuration. As shown in Figure 15b, we again observed strong correlations between the electrode and sound spectrograms, with sharp peaks at the specific frequencies of the pure sound stimuli (500, 1000 and 2000 Hz).

5.4.1.4 Audio-neural cross-correlations

In the previous sections, examples of highly correlated neural and audio recordings were shown. To further study the nature of this phenomenon, we used the cross-correlation to determine the delay between the two recordings that would maximize their correlation. For each channel and each frequency bin of the neural recording, we computed the correlation with different time lags applied to the audio signal at the same frequency bin. Figure 16 shows that most of the highest correlations are obtained for lags between -30 and 10 ms, in both speech production and sound perception cases. These very small lags suggest that the similar patterns occurring in the audio and neural signals are close to synchronous.

5.4.1.5 Objective assessment of acoustic contamination

The previous sections show that time-frequency patterns of audio signals are sometimes partially found in neural recordings. These correlations between audio and neural recordings occur at frequencies that correspond to the high-power frequency content of the sound (see sections 5.4.1.1-5.4.1.3) and seem to occur almost synchronously in both recordings (see section 5.4.1.4). These observations suggest a contamination of the electrophysiological measure by the audio signal through a physical phenomenon. This hypothesis of acoustic contamination is supported by the investigations on the origin of the phenomenon and its reproduction in section 5.4.1.3.

We developed an approach to assess the presence of contamination. We based this approach on the contamination matrix, which sums up the highest values in the audio-neural correlation matrices of a recording. The diagonal of the contamination matrix shows the audio-neural correlation for a given frequency while the other elements of the matrix represent cross-frequency correlations. Supposing that the contamination phenomenon is linear, we expect that power variations at a given frequency in the audio would cause power variations in the neural recording at the same frequency. Contamination is therefore characterized by high correlations limited to the diagonal. By contrast, when other sources of the electrophysiological signals (actual brain activity, muscle artifacts, motion artifacts) happen to be correlated with sound, the involved frequency bands are typically not exactly the same. In these cases, broad patches, vertical lines and/or horizontal lines of high values are observed in the contamination matrix. As detailed in section 5.3.4, the statistical criterion we propose compares the diagonal of the original contamination matrix to the diagonal of shuffled matrices. This allows to distinguish whether the high correlations are

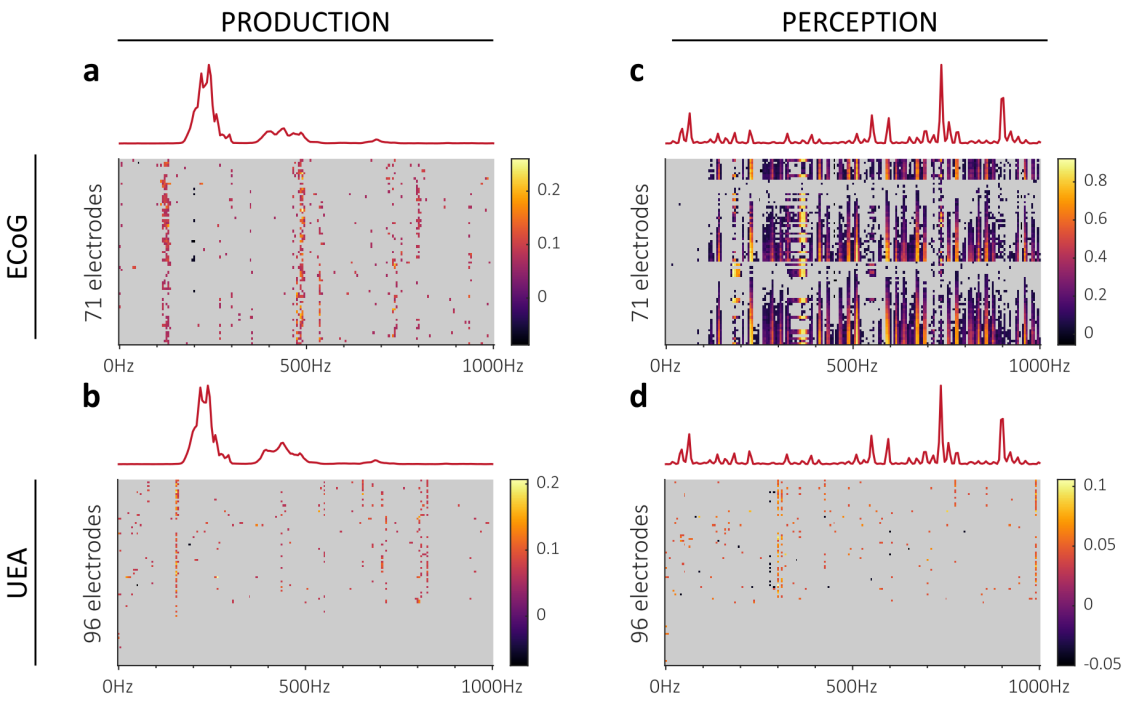


Figure 14. Spectrogram correlations between sound and neural data observed in the recording of participant P5 during session b. The results are presented depending on the experimental condition (speech production or perception) and the type of electrophysiological measure (ECoG and UEA). The heat maps show the value of the correlation coefficient with the spectrogram of the sound for each electrode and each frequency bin. The correlation coefficients that are not statistically significant are displayed in grey. The red curves indicate the mean PSD of the sound (a.u.) recorded during the experiments. The speech perception experiment used computer-generated vowels, peaks appearing in the mean PSD are the fundamental frequency and first harmonics of the vocal synthesizer. **(a, b)** Results for the speech production condition using ECoG and UEA data, respectively. **(c, d)** Results for the speech perception condition using ECoG and UEA data, respectively.

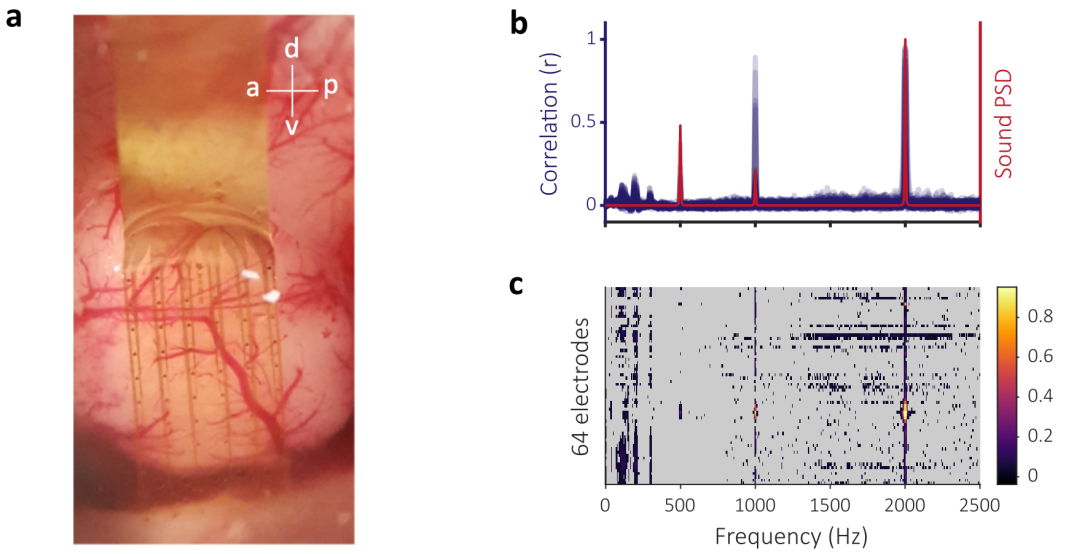


Figure 15. Correlations between sound and µ-ECoG spectrograms during pure tones perception in an anesthetized rat. (a) Photograph of a µ-ECoG grid positioned over the left auditory cortex of a rat. Directions: a = anterior, p = posterior, d = dorsal, v = ventral. **(b)** Each blue curve represents, at all frequency bins, the value of the correlation coefficients between the spectrogram of one electrode signal and the spectrogram of the audio signal. The red curve represents the mean PSD of the audio signal (a.u.). **(c)** Heat map of the correlation coefficient between audio and neural data for all electrodes and frequency bins. Correlation coefficients not statistically significant are displayed in grey.

1) limited to the diagonal, in which case they are significantly higher in the original matrix compared to the shuffled ones or 2) part of a larger patch of high correlations, in which case they are not significantly higher in the original matrix compared to the shuffled ones. We used this approach to evaluate 20 different speech production and sound perception datasets. Figure 17a shows the contamination matrices for 6 speech production datasets and Figure 17b the corresponding statistical evaluation by randomization. According to our statistical criterion P (see section 5.3.4), 4 datasets were contaminated (red vertical bars in Figure 17b) and 2 were not (green vertical bars). The audio-neural correlations that were observed for participants P2 (12) and P3 (Figure 13) can be observed on the diagonal of the contamination matrices (Figure 17a, top row). Very clear and highly statistically significant contaminations appear in these cases. Weaker but still visible and statistically significant contaminations also appear for P5 session b (ECoG) due to frequencies above 200 Hz and for D1 due to frequencies around 100 Hz. In recordings of participants P2, P3 and D1, lines of high correlation can also be observed outside the diagonal. They are due to the correlations between the voice's fundamental frequency and its harmonics.

Figure 18a shows the contamination matrices for 8 sound perception datasets, 5 of which are contaminated according to the corresponding statistical evaluation shown in Figure 17b. Contamination-specific patterns appear on the matrices for participants P5 (ECoG) and D2 (ECoG), for the p.m. session of participant CN (ECoG), for the rat

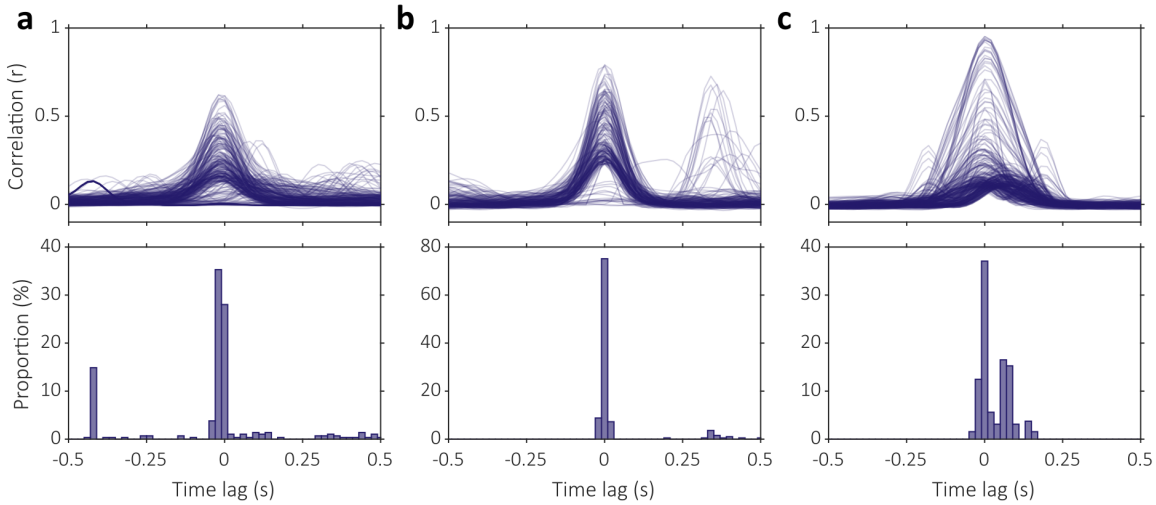


Figure 16. Audio-neural cross-correlations. Cross-correlations were computed with lags between -500 and +500 ms applied to the audio spectrogram (positive lags = delaying audio). All the frequency bins between 75 and 1000 Hz were considered. In the top graphs, each curve represents, for a given channel and a given frequency bin, the cross-correlation between the audio and the neural spectrogram. For each recording, only the 1% curves reaching the highest correlation values were kept. In each bottom graph, a histogram represents the distribution of time lags that maximize correlation for the curves in the corresponding top graphs. The histogram bins have a width of 20 ms. **(a)** Participant P2 ECoG recording during speech production. **(b)** Participant P3 UEA recording during speech production. **(c)** Rat micro-ECoG recording during sound perception.

recording (μ-ECoG) and for the in vitro experiment (ECoG). No contamination was found for participants P5 (UEA), CH1 (ECoG) and the a.m session of participant CN.

The full assessment of all datasets is summarized in Table 3.

5.4.2 Potential influence of contamination on speech decoding

We assessed the potential influence of contamination of electrophysiological signals by sound on the performance of neural decoding to predict acoustic features of produced speech. For this purpose, we considered ECoG data from participant P2 and P5 (session a). As can be seen in Figure 19 (panels a and b), the P2 recording was found to be contaminated in the 0-200 Hz range while the P5a recording was not. It should also be noted that the fundamental frequency of both participants exceeds 90 Hz and thus acoustic contamination cannot be observed below. The decoding performances of models using only neural features from 0 to 90 Hz were evaluated (Figure 19, panels c-d, top graphs). Models using neural features up to 200 Hz were then evaluated, and compared to the previous one in order to estimate the contribution of 90-200 Hz features (Figure 19, panels c and d, bottom graphs). We found that including neural features from 90 to 200 Hz resulted in an important increase in decoding performance for P2 and only a limited improvement of decoding accuracy for participant P5. This example shows that

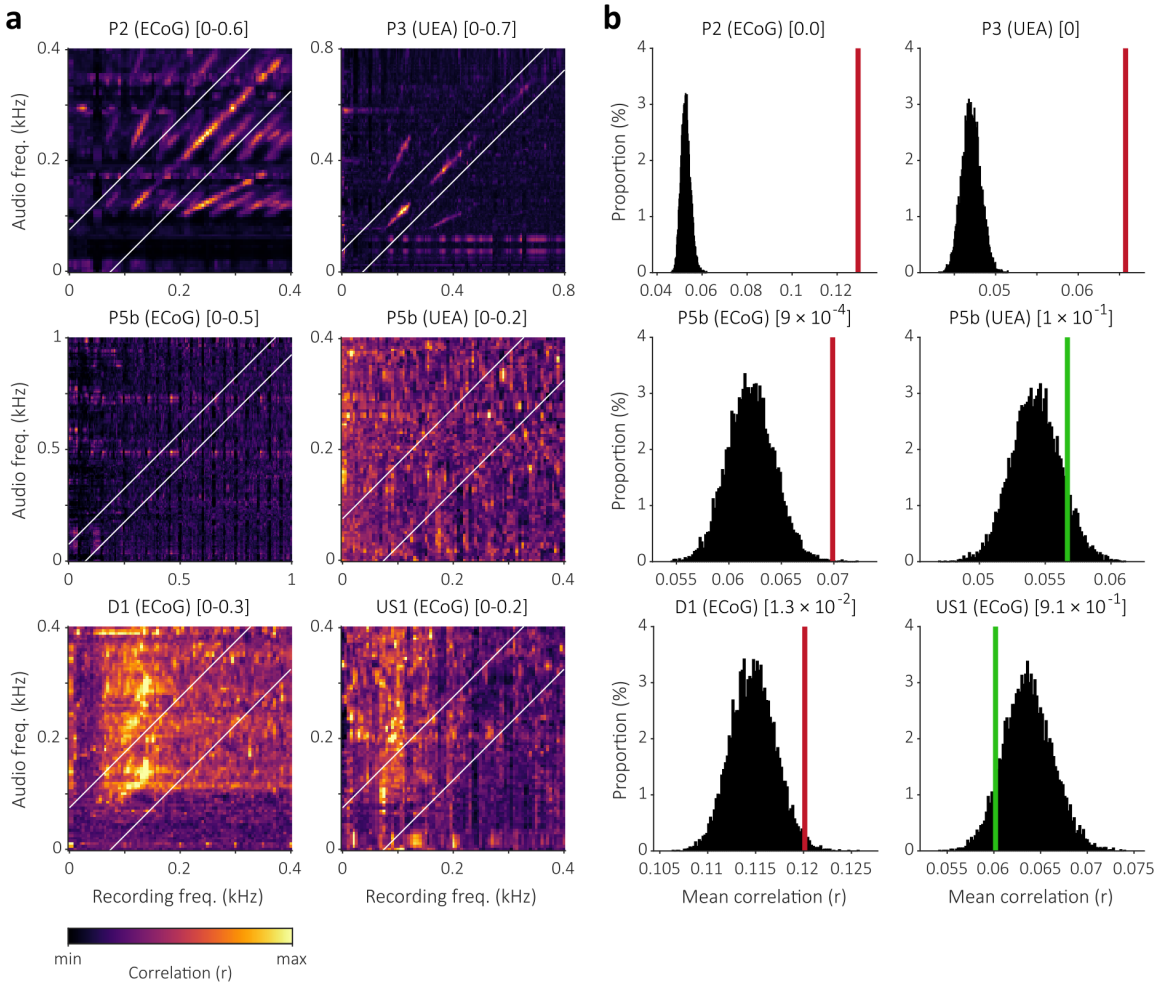


Figure 17. Objective assessment of acoustic contamination of 6 speech production datasets. (a) Audio-neural contamination matrices of each datasets. Each heatmap represents a recording. The color scale min and max values are indicated within brackets in the title of each heatmap (the max values were computed discarding frequencies below 75Hz, which do not contain any speech information, so that the colormaps were not influenced by 50-Hz or 60-Hz line noise). The white lines are a visual aid to assess the presence of high correlations on the diagonal. They are parallel to the diagonal of the matrix and cross the X and Y axes at 75 Hz. **(b)** Statistical assessment of contamination for the same datasets. The mean of the diagonal of the contamination matrix (vertical colored bar, red when statistically significant, green when not) is compared to the distribution of such values in 10000 shuffled contamination matrices (see section 5.3.4 for details). The estimated risk to wrongly consider the existence of contamination (P) is shown in square brackets for each dataset.

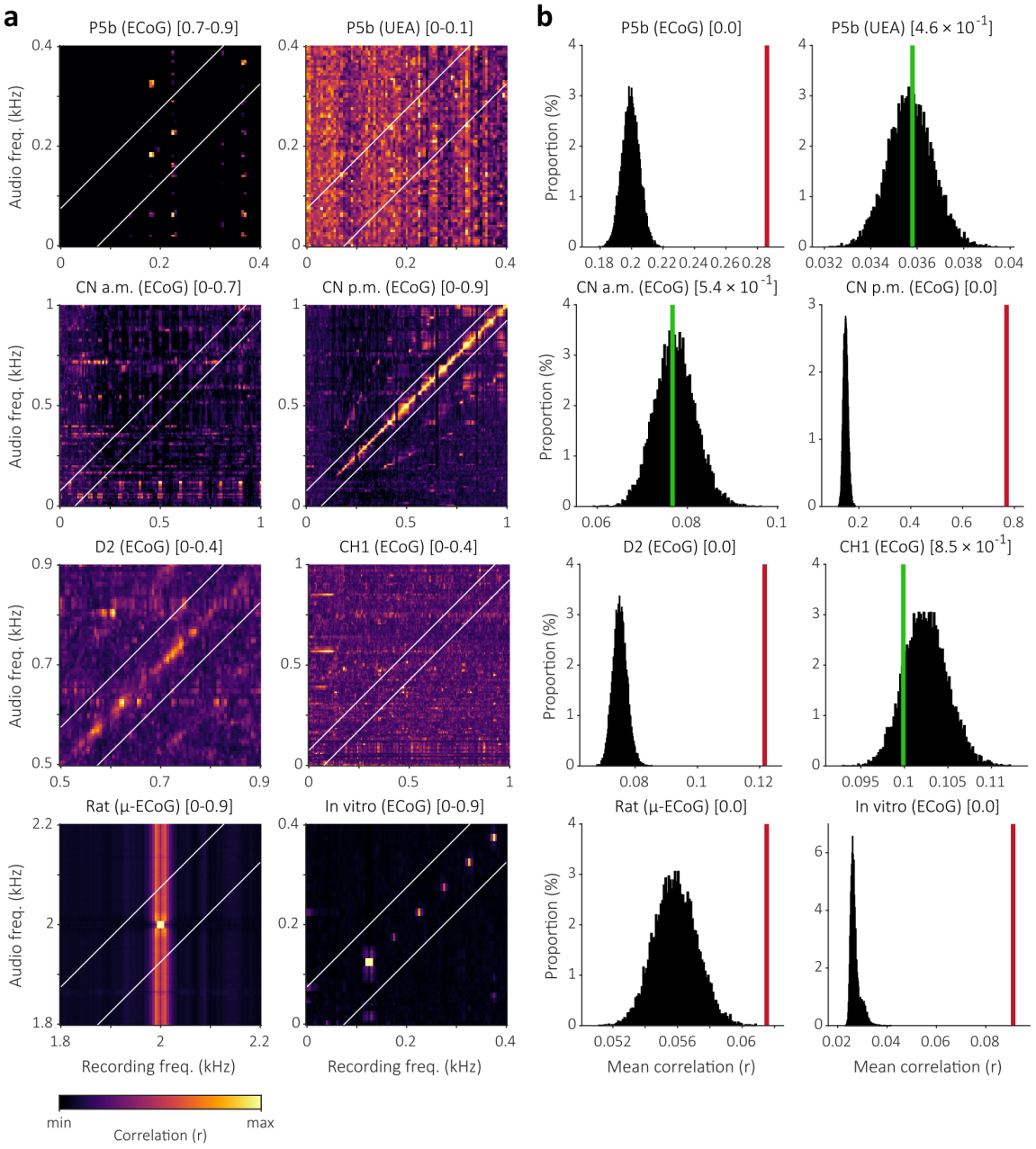


Figure 18. Objective assessment of acoustic contamination of 8 sound perception datasets. (a) Audio-neural contamination matrices of each datasets. Each heatmap represents a recording. The color scale min and max values are indicated within brackets in the title of each heatmap (the max values were computed discarding frequencies below 75Hz, which do not contain any speech information, so that the colormaps were not influenced by 50-Hz or 60-Hz line noise; for P5 ECoG we set the min value to 0.7 to increase visibility of diagonal correlations). The white lines are a visual aid to assess the presence of high correlations on the diagonal. They are parallel to the diagonal of the matrix and cross the X and Y axes at 75 Hz. **(b)** Statistical assessment of contamination for the same datasets. Same representation as in Figure 17.

Center	Participant	Electrodes	Task (session)	Duration (s)	Frequency range (Hz)	P
Grenoble	P2	ECoG	production	600	75-1000	$< 10^{-4}$ *
					0-200	$< 10^{-4}$ *
	P3	UEA	production	257	75-1000	$< 10^{-4}$ *
	P5	ECoG	production (a)	374	0-200	0.34
			production (b)	133	75-1000	7.0×10^{-4} *
			perception (b)	312	75-1000	$< 10^{-4}$ *
		UEA	production (b)	141	75-1000	0.11
			perception (b)	319	75-1000	0.46
	Rat	µ-ECoG	perception	600	75-2500	$< 10^{-4}$ *
Hangzhou	CN	ECoG	perception (a.m.)	40	75-1000	0.54
			perception (p.m.)	42	75-1000	$< 10^{-4}$ *
Frankfurt	D1	ECoG	production	98	75-1000	0.013 *
			perception	105	75-1000	0.94
	D1	ECoG	production	180	75-1000	0.46
			perception	151	75-1000	$< 10^{-4}$ *
Geneva	CH1	ECoG	production	109	75-1000	0.98
			perception	88	75-1000	0.85
	CH2	ECoG	perception	497	75-1000	0.99
Seattle	US1	ECoG	production	268	75-400	0.91
	US2	ECoG	production	133	75-400	0.46
	US3	ECoG	production	583	75-400	0.50

Table 3. Assessment of the presence or absence of contamination for 20 datasets from 5 different research centers. The durations reported refer to the time segments kept for analysis after the data selection step (detailed in section 5.3.1.2). For all datasets, the estimated risk to wrongly consider the existence of contamination (P, see section 5.3.4) is reported in the last column. Note that this value depends on the frequency range considered in the contamination matrix (indicated in the second to last column). Contaminated datasets are marked with an asterisk ($P < 0.05$). Their contamination matrices are presented in Figures 17, 18 and 19.

including a contaminated recording in a decoding study may positively bias the decoding performance and lead to overestimate the contribution of the high-gamma band.

5.4.3 Possible sources of acoustic contamination

5.4.3.1 Sound contamination and electrode quality

In the following of this paper, we investigate the possible causes of the contamination observed in neural signals. We first tested whether the level of sound contamination was determined by the quality of the electrode signal. Participant CN was recorded twice on a single day, once in the morning and once in the afternoon. Between the two sessions, the electrodes were disconnected and then reconnected after lunch. Sound contamination was observed only in the afternoon session (Figure 20a-b), which indicates that the contamination was not related in this case to the electrode array and its intracranial environment as those remained unchanged. Moreover, electrodes showing strong contamination in

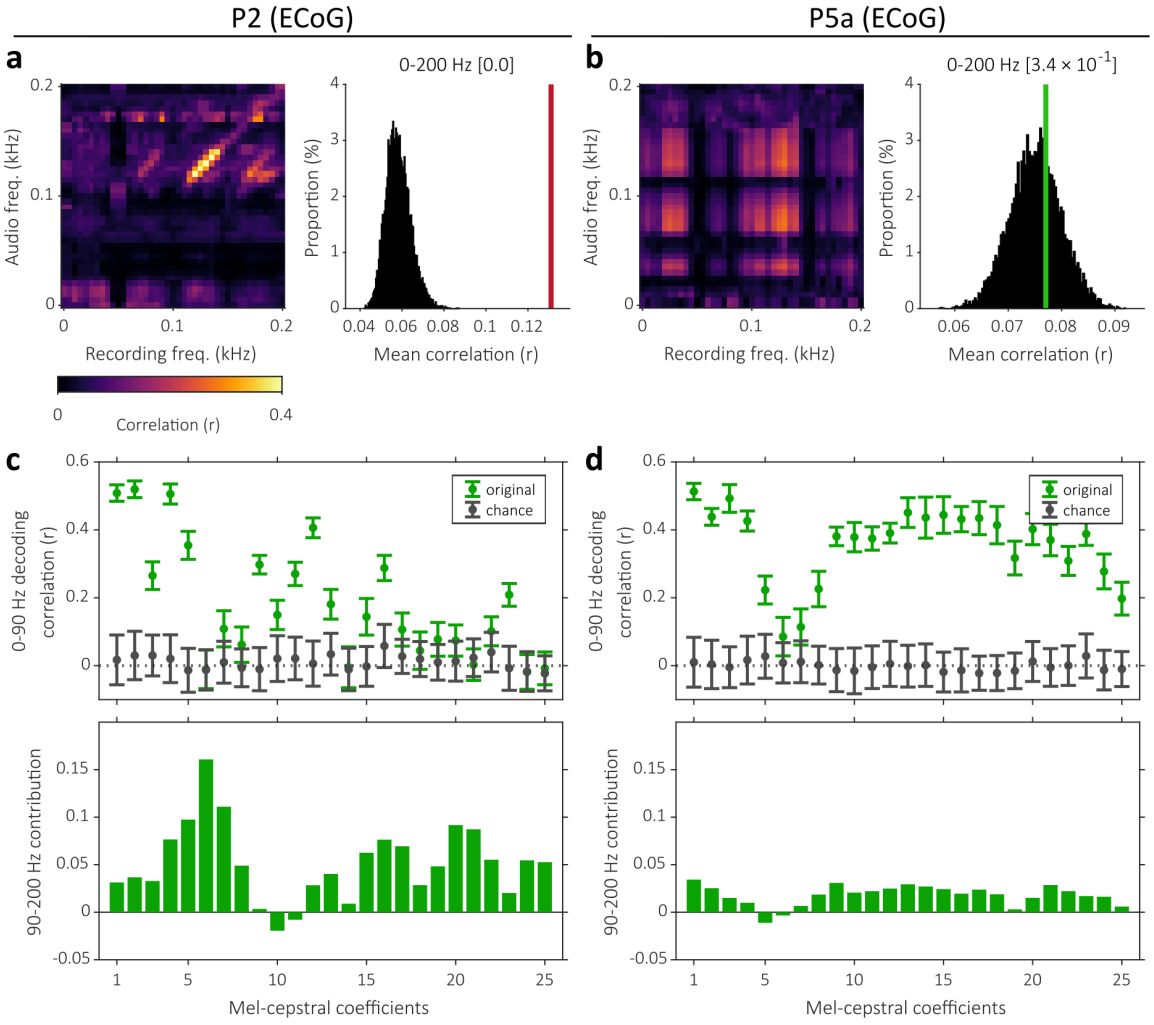


Figure 19. Linear decoding of speech features using either contaminated or not contaminated neural recordings. (a, b) Contamination matrix between 0 and 200 Hz (left) and corresponding statistical assessment of contamination (right) for ECoG datasets from participants P2 (panel a) and P5 (panel b), contamination P value is indicated between square brackets (P2 contaminated, P5 not contaminated). **(c, d)** Decoding performances for participants P2 (panel c) and P5 (panel d) of models using neural features from 0 to 90 Hz (top graphs) and the change in performance for models using neural features from 0 to 200 Hz (bottom graphs). The lower graphs can be interpreted as the contribution of features of the 90-200 Hz range.

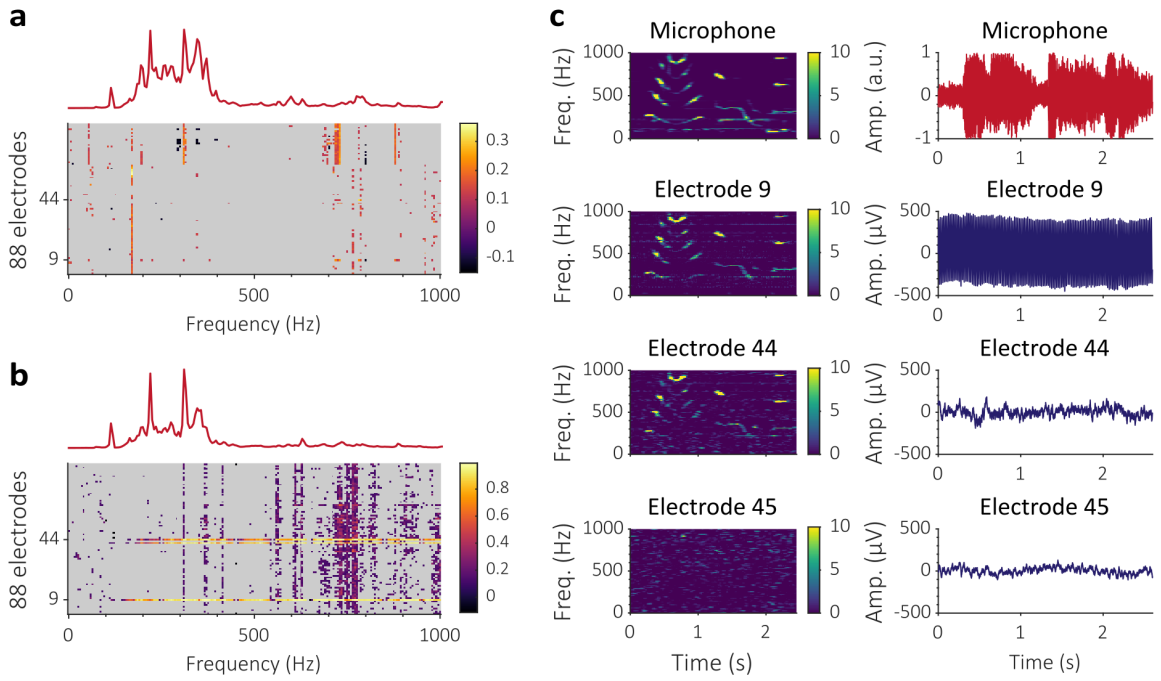


Figure 20. Correlations between sound and ECoG recordings in participant HG during speech perception. (a, b) Correlation heat maps (as in Figure 14) for the morning (panel a) and afternoon (panel b) sessions. **(c)** Example of 2.5-sec spectrograms (left) and raw signals (right) for the sound that was presented to the participant and signals from 3 electrodes of the grid, one with high 50-Hz power-line noise (electrode 9), and two with standard noise level (electrodes 44 and 45).

the afternoon session showed very variable signal quality (Figure 20c). For instance, one electrode with very strong 50-Hz power-line noise (considered as a typical "bad channel") showed a strong contamination, while two other channels with no such noise showed an equally strong contamination for one, and very weak or no contamination for the other. These observations indicate that the quality of the signal was not a sufficient predictor of sound contamination.

5.4.3.2 Electrode versus connector mapping of contamination

The study of how acoustic contamination affects the different channels of a recording in relation to their relative position can provide information about the physics of the phenomenon and its location along the acquisition chain. In most of the recordings available to us, the number of channels that were identified as contaminated was too low to observe spatial effects. However, in P5 ECoG recording during sound perception (session b) we observed that most of the channels were contaminated, with varying intensity (see Figure 14c). The audio-neural spectrogram correlation coefficients were averaged across all frequency bins to obtain a mean correlation coefficient for each channel. Then these values were mapped either on the grid layout or on the FEA connector layout. As shown

in Figure 21a-b, no clear spatial cluster of contamination was observed on the grid map. However, as could already be noticed in Figure 14c, it appears that channels 1-32 seem to show a continuous decrease of the contamination level, a pattern that is repeated on channels 32-64. Along the recording pipeline (described in section 5.2), the channels are grouped by 32 from the output of the splitter box to the input of the FEA. As shown in Figure 21c-d, we observed a very clear spatial organization according to the pin layout of the FEA input connector. The 4 connectors are interfaced with a single adaptor (Amplifier Manifold, Blackrock Microsystems, USA), which is fixed on the FEA case. The fact that higher contamination levels was consistently found on the top pins independently of the socket might be explained by a less tight fixation of the top of the adaptor, possibly causing the microphonic effect.

5.4.3.3 In vitro evidence of acoustic contamination

Next, we used a reduced experimental setup to determine more in details the cause of the observed correlations (see Figure 22). The experiment was designed to verify that the correlations between the sound and the electrode recordings can be obtained without brain activity and to attempt to demonstrate that the correlations originate from the mechanical transmission of sound vibrations. The electrical potentials of ECoG electrodes placed in PBS were recorded while pure tone sounds were played by the same loudspeaker as the one used to present sounds to Participant P5 and with similar intensity. In order to evaluate the intensity of the incident sound, a microphone was placed near the container filled with PBS. A soundproof box was used to insulate either the loudspeaker, the ECoG array, or part of the acquisition chain. The function of the box was to reduce the propagation of sound from the loudspeaker to the devices without substantially interfering with other parameters of the experiment. To determine the impact of sound propagation on the spectrogram correlations, we analyzed the data in open and closed box conditions.

In the first configuration, the loudspeaker was placed in the open box (Figure 22a). As for in vivo experiments, we found that high correlations occurred at some of the frequencies of the sound stimuli. For some electrodes, the value of the correlation coefficient at 125 Hz was larger than 0.9. This result demonstrates that spectrogram correlations similar to those described in sections 5.4.1-5.4.3 occur in absence of any brain activity. In the second configuration, the loudspeaker was placed in the closed box (Figure 22b). The reduction of the power of the incident sound due to the insulation is confirmed by the mean sound PSD (Figure 22b, top). We observed that most of the correlation coefficients also have much lower values (Figure 22b, bottom – compare with Figure 22a, bottom). This result supports the hypothesis of acoustic contamination, i.e. that the spectrogram correlations between sound and electrodes data originate from the mechanical propagation of sound to the neural recording hardware.

In the third and fourth configurations, the electrode array and the microphone were placed in the box but the rest of the acquisition chain was left outside. When the box was left opened (Figure 22c), we observed high correlations at the frequency of the stimuli, similarly to the previous open box condition (Figure 22a). The differences of frequency responses visible in the mean sound PSD across the 3 open box conditions can be explained by the modification of the arrangement of the experimental setup. In the last

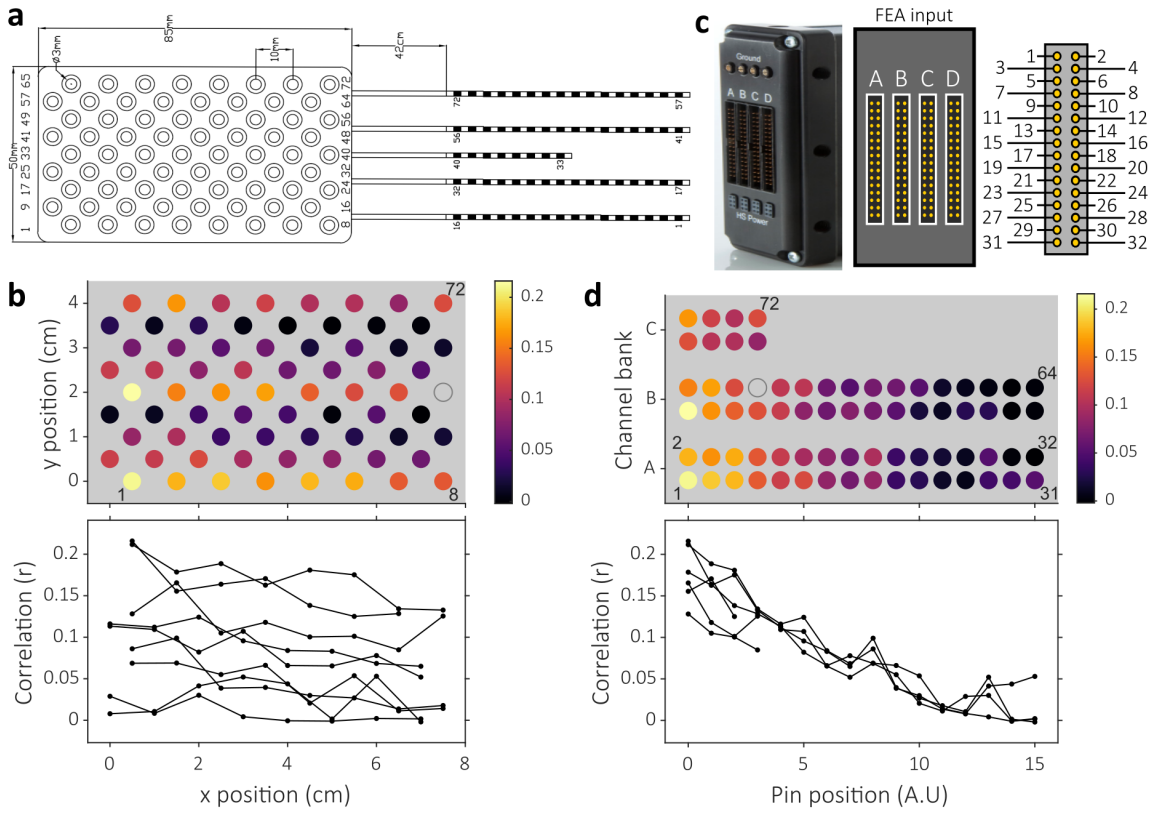


Figure 21. Spatial organization of acoustic contamination according to the grid and connector layouts. This anaylsi is based on the ECoG recording of participant P5 during sound perception). **(a)** Grid layout. **(b)** Top: The mean audio-neural correlation of each channel is represented as a colored dot at the position of the corresponding electrode on the grid (the empty circle represents a non-selected channel, see section 5.2). Bottom: Same values plotted along the rows of the grid. **(c)** Picture (left) and schematics (middle) of the front-end amplifier displaying four 32-channel connectors, each made of 16 pins (right). **(d)** Top: The mean audio-neural correlation of each channel is represented this time by a colored dot at the position of the connector pin connecting this channel (the connector layout is rotated 90° anticlockwise with respect to the picture and schematics of panel c). Bottom: Same values plotted along the rows of the connectors.

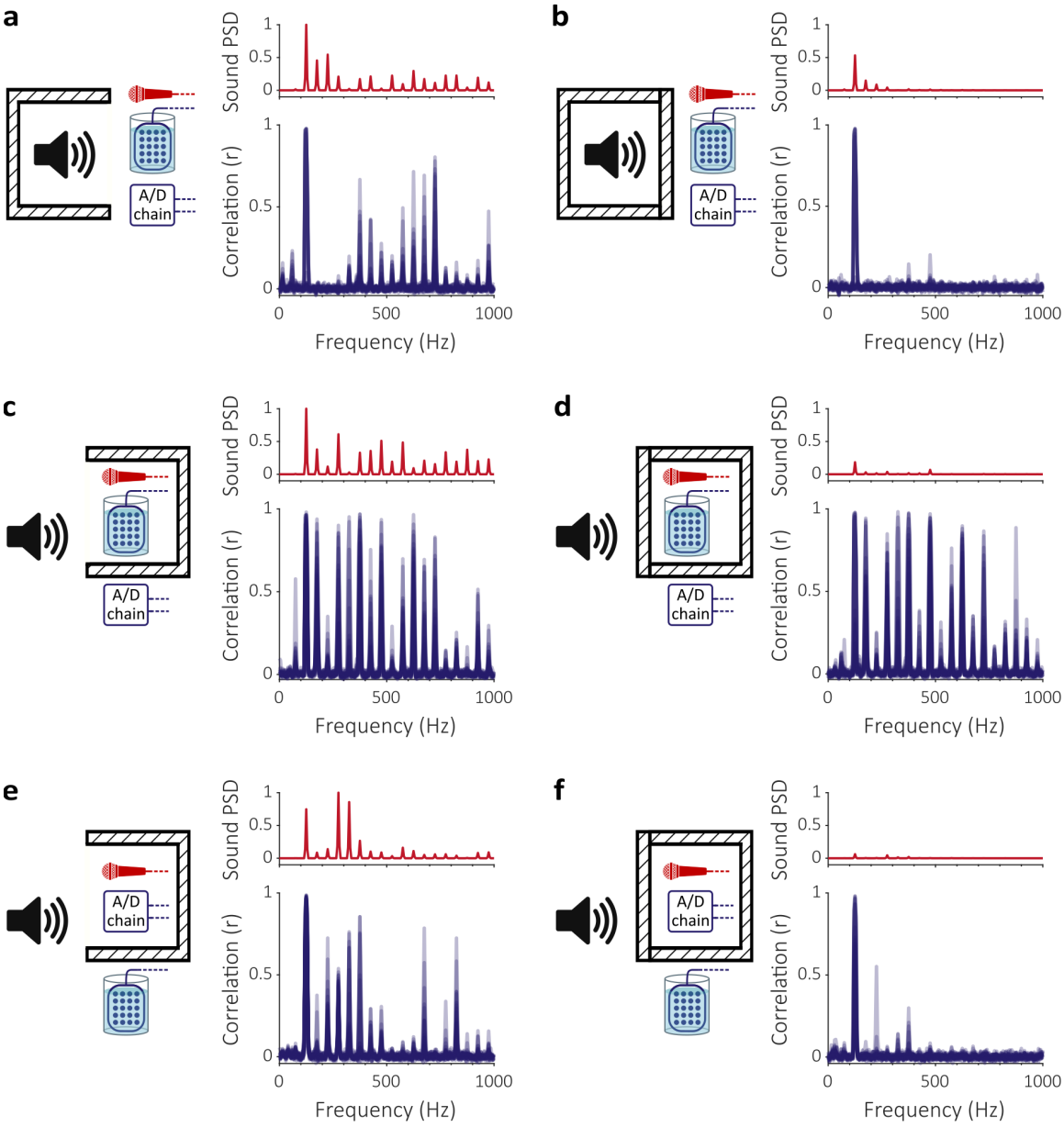


Figure 22. Correlations between sound and ECoG recordings in different in vitro experimental configurations. The red graphs show the mean PSD of the sound captured by the microphone. Mean PSD scale (a.u.) is common within each of the 3 panel pairs a-b), c-d) and e-f). In the blue graphs, each curve represents, at all frequency bins, the value of the correlation coefficients between the spectrogram of one electrode signal and the spectrogram of the audio signal. **(a)** Loudspeaker placed in the open box. **(b)** Loudspeaker placed in the closed box. **(c)** Electrodes and microphone placed in the open box. **(d)** Electrodes and microphone placed in the closed box. **(e)** Amplification and digitization chain (composed of the cables, adaptors, splitter box and FEA) and microphone placed in the open box. **(f)** Amplification and digitization chain and microphone placed in the closed box. Within each raw, only the fact that the lid was open or closed changed.

configuration, the box was closed over the electrodes and microphone (Figure 22d). The sound insulation provided by the box was confirmed by the large reduction of the sound stimuli mean PSD (Figure 22d, top). However, as shown in the bottom graph of 22d, the spectrogram correlations remained largely unaffected by the closing of the lid over the electrode array, contrarily to the previous experiment where the lid was closed over the loudspeaker (Figure 222b). This suggests that the acoustic contamination of the electrical potential measurement may not only occur at the electrode level but also at other levels of the acquisition chain. To test this hypothesis, the amplification and digitization chain (A/D chain, composed by the cables, adaptors, splitter box and FEA) was put inside the sound-attenuating box with the microphone. In this case the electrodes in PBS were outside the box. While correlations were high when the box was open (Figure 22e), they were strongly reduced when the box was closed (Figure 22f). This further confirmed that the acoustic contamination mainly occurs in the recording chain and not at the electrodes level.

5.4.3.4 Localization of acoustic contamination along the recording chain

Finally, we aimed at determining where along the recording chain the contamination occurred. The fact that contamination was observed in participant CN in the afternoon but not in the morning session suggests that disconnecting and reconnecting the electrodes to the system could have produced the contamination to occur in the afternoon. To test this more thoroughly, sounds were delivered very locally at different locations along the recording chain connected to electrodes bathed in PBS (Figure 23). The statistical criterion P indicated that contamination was found for every location of the sound delivery but with varying intensity, as can be seen on the channel-frequency correlation heat maps (Figure 23c). Only very weak contamination could be observed when the sound was delivered next to the electrodes in the PBS solution. By contrast, a clear contamination was observed when the sound was delivered against the ECoG grid cables, the pigtail connectors, or the splitter box touchproof connectors. This result is coherent with the idea that contamination is caused by the mechanical vibration of hardware elements, as observed in section 5.4.3.2 at the level of the FEA input connector in the case of P5 ECoG perception recordings.

5.5 Discussion

Data considered in this study includes human and animal recordings during speech production and/or sound perception tasks. Using these different setup conditions, we observed statistically significant correlations between the spectrograms of electrophysiological and simultaneously recorded audio signals. These correlations occurred at frequencies with the highest spectral powers in the sound signal, thus encompassing the high-gamma range and also frequencies above 300 Hz. This contamination effect was observed in recordings from ECoG and μ-ECoG grids and intracortical micro-electrode arrays, interfaced with different data acquisition systems. The phenomenon was observed in data collected by 3 out of the 5 centers worldwide who participated in this study. Thus, this variety of

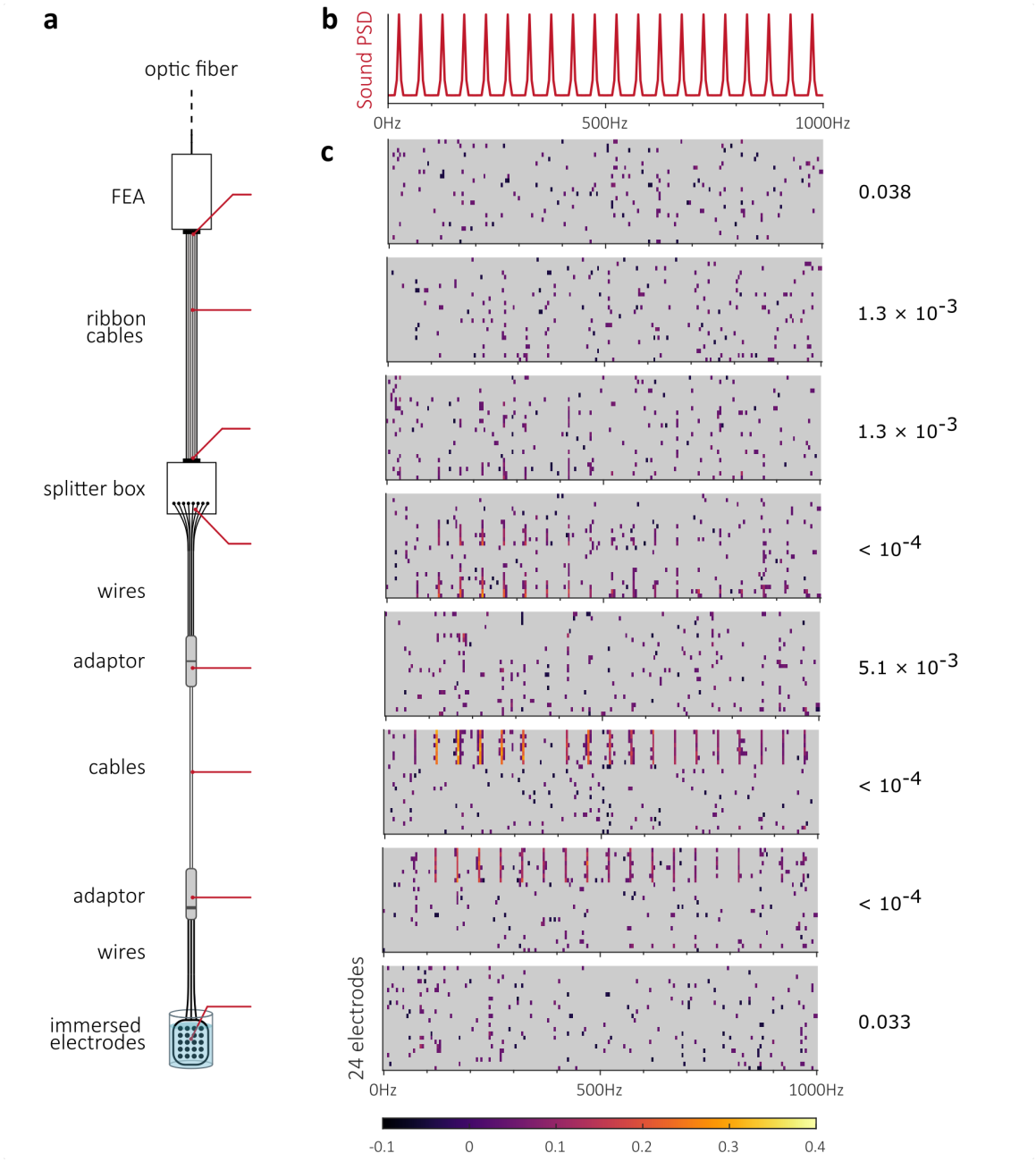


Figure 23. Determination of the location of sound contamination along the recording chain (using ECoG electrodes). (a) Sounds were delivered focally at different locations (indicated by the red lines) of the recording chain. **(b)** Mean PSD of the sounds delivered through the speaker. **(c)** Correlation heat maps for each location of sound delivery (each map is displayed against the corresponding location of sound delivery indicated in panel a). The estimated risk to wrongly consider the existence of contamination (P) is indicated on the right of each heat map.

recording conditions suggests that acoustic contamination of neural signals is a widespread phenomenon.

We observed sound contamination of neural signals in different setups. We could reproduce the phenomenon in a minimal in vitro setup, confirming that sound-electrode correlations do not originate from brain activity and arise from the impact of sound vibrations on the acquisition chain. The experiments shown in sections 5.4.3.1-5.4.3.4 further suggest that in the tested setup, the microphonic effect does not necessarily take place at the electrodes' level, but in the rest of the recording chain. In section 5.4.3.1, two recordings involving the same participant on the same day show different levels of contamination, which can be attributed to the disconnection of the recording hardware between the two sessions. Analyses of the spatial distribution of contamination in a highly contaminated recording showed that it was coherent with a microphonic effect occurring at the input connection of the amplifier but not at the electrode level (see section 5.4.3.2). In vitro experiments show that isolating the acquisition chain from sound reduces the contamination, as opposed to isolating the electrodes (see section 5.4.3.3). Focal sound delivery at different locations along the chain showed the acoustic contamination was prominent mainly at the level of cables and connectors. Improving elements composing the recording chain thus appears mandatory to ensure proper and artifact-free neural signal recordings.

Although contamination could occur through a crosstalk between channels within the same hardware acquiring simultaneously sound and neural signals, we excluded such possibility in several of the recordings considered here. In particular contamination was observed in a rat recording while in these the sound and the neural signals were acquired with two separate hardware (CED micro1401 and Intan systems). We observed that there is actually no particular hardware element responsible for the contamination. Rather, the quality of interconnection is critical and should be verified systematically. For instance, when participant P5 was presented with synthetic speech through a loudspeaker, the microphonic effect likely stemmed from the quality of the FEA connector (Figure 21). However, for in vitro PBS recordings, the contamination was small at the level of the FEA connector and more important at the level of the headbox and cables situated upstream (Figure 22). Moreover, the same setup used with different participants (or even the same participant in two different sessions) may sometimes exhibit contamination and sometimes not (as here with participant CN). Thus, a given setup should likely not be granted for clean once and for all. Rather, we suggest that any dataset should be objectively tested for any contamination before being considered for further analysis.

The extent to which the acoustic noise spectrally overlaps with the measured brain activity depends on the nature of the sound and on the studied activity. In the case of ECoG recordings during speech production paradigms (see section 5.4.1), the overlap between the range of the voice fundamental frequency and the high-gamma band might compromise recording artifact-free signals in this band. As suggested by results in section 5.4.3, sound stimuli, and by extension any sound during the recording, could contaminate the recorded data in any frequency band. In particular, high-frequency components of the sound may also influence the detection of multi-unit activity in micro-electrode recordings (see section 5.4.2). This is all the more important that several studies have reported best decoding performance when using a window for the neural features centered with respect to the current time point of the speech feature to be decoded (Martin et al.,

2014; Chartier et al., 2018; Anumanchipalli et al., 2019; Herff et al., 2019). In such case, contamination occurring at delays inferior to 10 ms (Figure 16) would bias the decoding results.

While these results demonstrate what could be seen as a relatively trivial contamination of electrophysiological recordings by surrounding acoustic signals, the implication of the study is important in both the neuroscience and neuro-engineering domains. In particular, the common investigation performed here on several datasets acquired in various research places worldwide (France, China, Germany, Switzerland, and the United States) suggest that decoding analyses should be performed after having excluded any potential microphonic effect.

Yet and importantly, this report does not question the existence of relevant physiological neural information in high-gamma frequency signals underlying speech production or sound perception. Several groups have shown that spectral features of imagined speech or silent articulation can be predicted to some extend from low or high-gamma signals recorded in participants that are not overtly speaking (Pei et al., 2011a; Ikeda et al., 2014; Martin et al., 2014; Bocquelet et al., 2016a; Martin et al., 2016; Anumanchipalli et al., 2019; Gehrig et al., 2019). Also, contamination by mainly the fundamental frequency is insufficient to explain the decoding performance of sublexical features such as articulatory gestures and phonemes, especially consonants (Chartier et al., 2018; Mugler et al., 2018). However, we think it is important for past and future studies assessing the contribution of high-gamma or multiunit activity to speech decoding to make sure that neural signals are free of acoustic contamination in the considered frequency bands.

The purpose of this study is therefore to alert on possible microphonic contamination of neural signals, especially when building decoders of neural activity underlying overt speech production or sound perception. Future developments of speech prostheses should thus build upon these findings. In particular, experimental setups should be improved to become less sensitive to microphonic effects, and signal-processing techniques should be developed to eliminate sound contamination in neural recordings. Meanwhile, data should be carefully tested ahead of further decoding analysis.

Chapter 6

Dynamics of neural activity in Broca's area during overt and covert speech

6.1 Introduction

Given its key role in speech production, Broca's area (BA) could be an interesting recording area for speech BCIs. It has been shown in several studies that ECoG recordings of BA could be used to decode speech features Pei et al. (2011a); Mugler et al. (2018); Anumanchipalli et al. (2019). However, little is known on the potential of the neuronal activity in BA for the decoding of speech. Understanding the activity of a neuronal population in BA during overt and covert production could therefore lead to interesting developments in future speech decoding studies.

As discussed in section 3.1.2.1, Broca's area (BA) plays a critical role in speech production but its precise functions remain unclear. In several studies, BA was claimed to be mostly involved in the stages preceding or initiating the generation of articulatory commands. In particular, it was hypothesized that BA acts as an intermediary between acoustic targets and articulatory representations (Guenther and Vladusich, 2012; Flinker et al., 2015). This view is coherent with the idea that BA does not actually manipulate articulatory representations but phonological content (Indefrey and Levelt, 2004; Tate et al., 2014). Several studies on speech processing have also attributed higher language-related functions to this area, including semantic and lexical processing (Sahin et al., 2009; Hagoort and Indefrey, 2014; Goucha and Friederici, 2015).

The activity of BA during overt and covert speech has been compared in different studies. It was found using fMRI (Huang et al., 2002; Palmer et al., 2001) and ECoG (Pei et al., 2011b) that BA is active during covert speech production, but with lower activation levels that for overt speech. It remains however unknown if the dynamics of the neuronal activity in BA are similar during both conditions.

The current findings and hypotheses about Broca's area have, to our knowledge, never been studied at the level of individual neurons. In the present chapter, the dynamics of a population of neurons in Broca's area during overt and covert speech are analyzed based on the intracortical recording of one participant. Firstly, we expose how the firing rates of these neurons are modulated during the participant's overt and covert speech. In a second

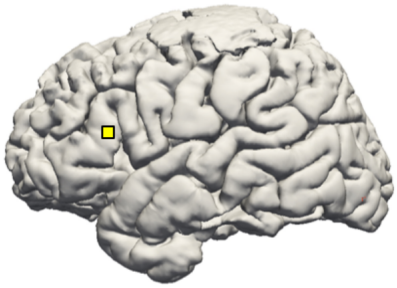


Figure 24. Localization of the Utah array on the reconstruction of the brain of participant P3. The reconstruction was obtained from pre-operative MRI.

part, we show that spiking activity in BA relates to the phonological content of overtly pronounced sentences. Finally, we report speech-related low frequency LFP activity and examine their relation with the spiking activity.

6.2 Material and methods

6.2.1 Micro-electrode recording

The dataset consists of a peroperative recording using a 96-channel Utah array (Figure 2c) implanted in Broca's area (pars triangularis), as represented in Figure 24. Audio was recorded synchronously with neural activity. The data acquisition was described in details in section 5.2.

6.2.2 Protocol

The data was acquired in successive trials. In each trial, the participant was asked to read, repeat and covertly repeat a French sentence displayed on a screen. Some of the sentences were sequences of two or three vowels. The trials were driven by cues displayed on a screen as shown on Figure 25. The participant was asked to say "OK" to signify the end of the covert repetition. Out of the 43 recorded trials, only the first 31 were analyzed, for reasons detailed in section 6.2.4.2.

The chosen sentences were part of the articulatory-acoustic corpus BY2014 (see section 7.2.3). The corpus of sentences in P3 experiment contained a majority of verbal sentences, some nominal sentences and four sequences of vowels (for example "a, i, ou"). For each trial, one sentence was read out loud, repeated out loud and then covertly repeated (see figure 25). Each of the succession of vowels appeared in two trials, all other sentences appeared only in a single trial. The 31 selected trials contain 27 different sentences, 62 speech epochs and 31 covert speech epochs.

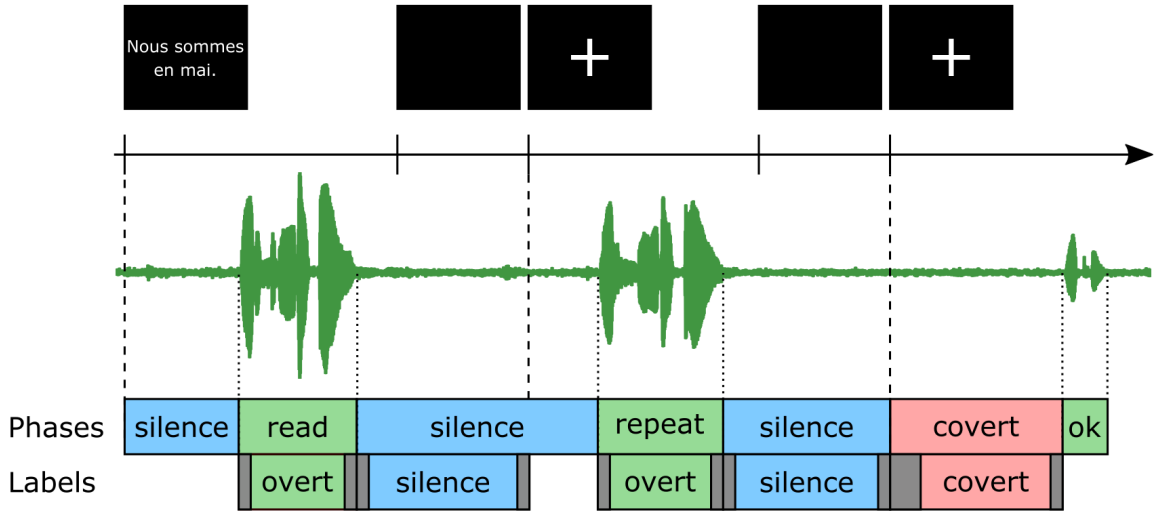


Figure 25. Course of a trial for the standard speech production experiments performed by participant P3. The black rectangles represent the visual cues that were presented to the participant. The timeline indicate the time of appearance of the visual cues. The green curve represents the audio signal. The first line of colored boxes indicate how the trials were divided into phases corresponding to the different cognitive tasks performed by the participant. The second line indicate the labeled intervals that were considered for the classification of neuronal activity. On that line, the grey boxes show the margins that were excluded from each interval.

6.2.3 Trial phases labeling

The start and end times of utterances where manually annotated using the audio data. As illustrated in Figure 25, each trial was broken down into several phases using the annotated times as well as the times corresponding to stimulus display:

- the "read", "repeat" and "OK" speech phases, delimited using audio
- the covert phase, delimited by the instruction stimulus and the "OK"
- the silence phases

When averaging the results from different analyses across trials in Figures 30, 31c and 36a, the boundaries between each phase were aligned so as to preserve the phase-related modulations. Before averaging, the data was transformed so that phases have the same duration in each trial. For each phase (for example overt reading), the average duration across trials was computed. In each trial, each was temporally stretched or compressed so that its duration matches the average duration. Signals were temporally stretched on compressed using linear interpolation.

6.2.4 Neural signal processing

6.2.4.1 Common average reference

The problem of common average reference (CAR; see section 5.3.1.6) is that it can introduce channel-specific noise, in particular high-amplitude transient artifacts, to the other channels. In order to mitigate this issue, it is possible to replace the average of all channels by their median. The use of median instead of mean for CAR has been shown to increase the amount of task-related information in ECoG recordings (Liu et al., 2015) and to increase the number of detected spikes in intracortical micro-electrode recordings (Rolston et al., 2009). Here, median CAR was applied to the neural signals.

6.2.4.2 Spike-sorting

Prior to spike-sorting, band-pass filtering was applied to the neural data. The band-pass filter was designed to select frequency content between 400 and 3000 Hz. The neural signal of each channel was set to zero mean and scaled to have an MAD-based standard deviation of one. As discussed in section 5.3.1.4, using the MAD to estimate the standard deviation makes it more robust to outliers. In the case of spike detection, it allows to estimate the standard deviation of the signal without spikes (Martínez and Quiroga, 2013). The spike sorting process was carried out using Offline Sorter (Plexon, USA). The spike detection thresholds were set to plus and minus 5 times the MAD-based standard deviation of each channel. Waveforms were extracted by selecting the 2 ms window around the detected peaks (0.7 ms before the peak and 1.3 ms after). The detected waveforms were projected in a feature space using PCA. Spike-sorting was then performed in a semi-automatic way. First, the T-Distribution expectation-maximization algorithm was applied. This algorithm fits a mixture of T-Distributions to the point densities in feature space by adjusting the number of distributions and their parameters. Finally, waveforms and clusters were manually adjusted. 64 putative neurons (or units) were obtained from 31 channels. In the following, each unit is designated using a unique 4-digit number where the 2 first digits refer to the channel on which they were recorded. Most of the units were not detected anymore after trial 31, probably due to a relative movement between the implant and the brain. For this reason, only the first 31 trials were analyzed in the following steps. An illustration of the spiking data is visible in Figure 26.

6.2.4.3 Firing rates

The average firing rates during the selected trials ranged from 0.02 spikes/s for the slowest-firing unit to 7.4 spikes/s for the most active one (with an average of 1.3 spikes/s). Smoothed firing rates were computed by convoluting the spike trains with a Gaussian kernel (Nawrot et al., 1999). This method, relying on the hypothesis of rate coding, was developed as an alternative to trial-averaged peri-stimulus time histograms (PSTHs). It is useful to study the activity of single neurons on single trials, especially when the activity is not precisely time-locked to external events. A Gaussian kernel with a standard deviation of 300 ms was chosen for the smoothing. This kernel was wide compared to the ones used in other studies, which is coherent with the comparatively low firing rates in our dataset.

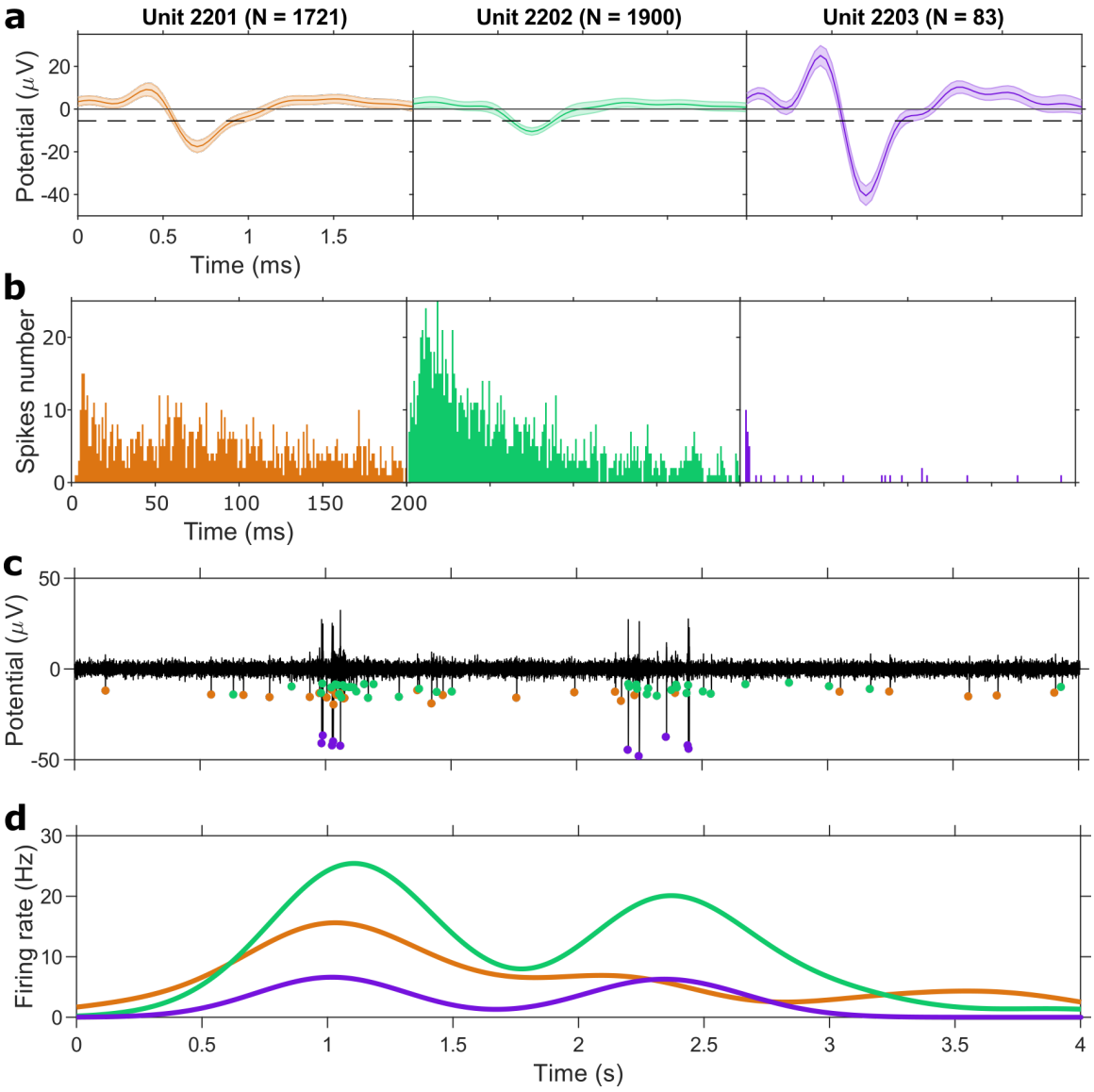


Figure 26. Illustration of the spike-sorting output. (a) Average waveforms of the three units detected on channel 22. The shaded area represents the standard deviation. The dotted line represents the negative threshold. N indicates the number of spikes detected for each unit. **(b)** Distributions of the inter-spike intervals for same 3 units as in panel a. **(c)** Extract of filtered signal from channel 22. The colored dots indicate detected spikes corresponding to the units of the upper panel. **(d)** Smoothed firing rates corresponding to extract displayed on panel b.

6.2.4.4 Slow LFPs

The slow variations of the local field potentials were extracted using low-pass filtering. The low-pass filter was designed to select frequency content inferior to 1 Hz. The LFP variations were then downsampled to 50 Hz.

6.2.5 Condition classification

In order to assess the difference of firing rates between overt speech, silence and covert speech, classification of those 3 conditions was carried out. The performance of the classifier was assessed by leave-one-out cross-validation on the 31 trials.

6.2.5.1 Labeling

As illustrated in Figure 25, labeled epochs corresponding to overt speech, silence and covert were defined in each trial. The labeled epochs were defined as subparts of the trial phases. The periods preceding overt speech were excluded from classification in order not to risk to include neural activity related to speech preparation and stimulus presentation in the silence data. In order not to include samples for which labeling was ambiguous, 300-ms margins at the beginning and end of each phase were also excluded from the classification. At the beginning of covert speech epochs, a 600-ms margin was removed instead, in order to take into account the participant's reaction time before actual covert speech. 600 ms indeed corresponds to the rounded value of the average time reaction in the repetition task. To evaluate the classification process, chance levels were estimated by randomly shuffling the labels of these labeled epochs.

6.2.5.2 Features

The smoothed firing rates sampled at 20 Hz were used as features for the classification. Each observation was composed of the firing rates of the 64 units with added temporal context. The context covered the 300 ms preceding and following the observation time, resulting in a neural features space with $64 \times 13 = 832$ dimensions.

In each fold, a subset of features were selected using Kruskal–Wallis one-way analysis of variance. For each feature, this test returned the p-value for the null hypothesis that the feature's distribution was identical for the 3 classes. The features with a p-value inferior to 0.01 were selected.

6.2.5.3 Linear discriminant analysis

The classification was carried out using linear discriminant analysis (LDA). This model was chosen for its simplicity, as is has a no hyperparameters to tune and is not very prone to overfitting due to its low number of internal parameters. LDA fits multivariate Gaussian densities to each class. It assumes that the feature data are normally distributed and that the covariance matrices of the classes's Gaussian distribution are equal. During training, the mean of the Gaussian distributions and their common covariance matrix are estimated so as to minimize a misclassification cost.

When applied to an observation, the trained LDA model returns for each class the posterior probability that the observation belongs to it. This posterior probability is a product of the prior probability and the modeled multivariate normal density. Here the prior probability, the probability for an observation to belong to a class independently of its value, was considered uniform – that is to say equal to $1/3$ for all 3 classes.

6.2.5.4 Mathews correlation coefficient

Additionally to the confusion matrix, the classification results were evaluated in a more concise way using the Matthews correlation coefficient (MCC). The MCC is an evaluation metric for binary classification that is computed using the following formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In this formula, TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. The MCC returns the same values as Pearson's correlation coefficient computed between the actual and predicted labels. It returns $+1$ for a perfect classification and -1 if the prediction is the exact opposite of actual labels. MCC was chosen because it has advantages over the commonly used metrics of accuracy and F1-score, especially on imbalanced datasets (Chicco, 2017; Chicco and Jurman, 2020). It was used to evaluate the classification of each class against the two others.

6.2.5.5 Phoneme segmentation

The sentences of the corpus were segmented into phonemes based on the audio recording. This process used hidden Markov models on MFCCs (seen section 2.1.3.2) and was implemented using the HTK toolkit[1].

6.2.6 Spike train analysis

6.2.6.1 Spike train distance

In order to estimate the similarity between spike trains, we used the Victor and Purpura distance (VP distance) based on spike times (Victor and Purpura, 1996, 1997). This metric is computed as the cost of transforming a spike train S_a into a spike train S_b. This transformation is accomplished through a succession of elementary steps whose individual costs are added together to form the total cost (see Figure 27). The distance between S_a and S_b is given by the smallest total cost of any sequence of elementary steps that transform S_a into S_b. The available elementary steps are: adding a spike, deleting a spike and shifting a spike in time. Both adding and deleting a spike are assigned a cost of 1. The cost of shifting a spike is controlled by the parameter q so that shifting the time of occurrence of a spike by Δt is assigned the cost $q|\Delta t|$. In the absence of spikes during one or both of the epochs being compared, the distance was not included in further analyses.

[1] Available at https://htk.eng.cam.ac.uk/.

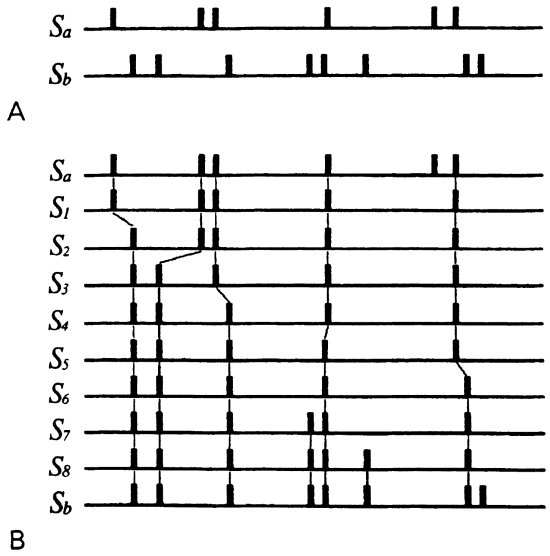


Figure 27. Illustration of the VP distance. (A) Spike trains S_a and S_b whose distance is to be determined. **(B)** S_i $(i = 1, 2, ...8)$ represent a path of elementary steps transforming S_a in S_b. Reproduced from Victor and Purpura (1996).

If $q = 0$, the distance between two spike trains is equal to the difference in the number of spikes. If $q = \infty$, the distance between two spike trains (that are not perfectly synchronized) is equal to the sum of the number spikes. For intermediate values, q, measured in s^{-1}, expresses the sensitivity of the metric to the timing of spikes. When comparing two spike trains S_a and S_b containing only one spike each, the two candidate sequences of elementary operations to transform S_a into S_b are: 1) shifting the spike of S_a and 2) deleting and inserting a spike in S_a. Shifting the spike will result in a lower cost only if the two spikes are separated by less than $2/q$.

Given the temporal variations between two repetitions of the same sentence, we assumed that associated spike trains could not be expected to precisely match. Moreover, given the very low firing rates in the considered epochs, we hypothesized that this number of spikes might be more related to the content of the sentences than their precise timing of occurrence. q was therefore set to $1\,s^{-1}$ so that $2/q = 2\,s$. Since most of the speech epochs last less than 2 s (97%), this choice for q implies that shifting a spike from the beginning to the end of the train costs less than deleting or inserting one. As a result, the spike trains with the same number of spikes tend to be considered less distant than those with a better temporal match on a part of the spikes but different number of spikes. In other words, the chosen metric takes into account temporal coding but is primarily influenced by rate coding.

In order to have an indicator of the spike trains distance that takes all units into account, a multi-unit distance was defined by averaging all unit-specific distances. Since the spike train distance tend to be larger for units with higher average firing rate, the unit-specific distances were normalized to zero-mean and unit-variance before taking the average. The multi-unit distance between the spike trains corresponding to epochs e_i and

e_j is noted $D_{spikes}(e_i, e_j)$.

6.2.6.2 Spike train classification

The spike trains occurring during overt speech were classified with respect to 27 classes, corresponding to the uttered sentences. The classification used the multi-unit spike train distance to perform basic pattern matching in a leave-one-out fashion. In each fold, the multi-unit spike train of one of the 62 utterances was isolated. The classification "model" consisted in the 61 remaining spike trains labeled with the 27 classes. D_{spikes} was computed between the spike train to be classified and the spike trains of the "model". A score was computed for each class by averaging the distances obtained with spike trains of this class. Finally, the test spike train was assigned the sentence label corresponding to the lowest score. The chance accuracy was estimated 1000 times by shuffling the sentence labels and repeating the whole process. The proportion of chance accuracy values exceeding the true accuracy was used to assess the significance.

6.2.7 Sentences similarity

One way to quantify the similarity of a pair of speech epochs is to categorize them as identical, when they contain the same sentence, or different, when they contain two different sentences. Here we also introduce two ways of estimating the distance between speech epochs, one based on their phonemic content and another based on their semantic content.

6.2.7.1 Phonemic distance

The BY2014 corpus features the phonemic transcription of each sentence. For example, the transcription of the sentence "Une nuit bleu horizon" is /y n ə n y i b l ø o ʁ i z ɔ̃/. In order to quantify, the phonemic distance between two sentences, the Damerau-Levenshtein string-edit distance was used (Damerau, 1964; Levenshtein, 1966). It counts the minimum number of elementary operations required to transform one sequence of characters to another, the elementary operations being: adding or deleting a character, replacing a character by another one or swapping adjacent characters. The phonemic distance will be noted $D_{phonemic}$ in the following.

6.2.7.2 Semantic distance

In order to quantify the semantic similarity of sentences, we mapped the words composing the sentences to vectors using the concept of word embedding. The idea of word embedding is to obtain a low-dimensional representation of words, so that the representations of words with similar meanings are close in the vector space (see section 2.2 for more information). Here the word representations produced by a skip-gram word2vec model was used (Mikolov et al., 2013). The model was trained on the French CoNLL17 corpus (5494845765 tokens, 2567698 unique words) and used the 10 preceding and following words as context. The trained model was used to produce 100-dimensional vectors for

each of the unique words in the corpus. These vectors were made freely available by the Nordic Language Processing Laboratory[1]. Using this database, we could associate a word vector to each of the words in the sentences pronounced by participant P3.

In order to retain information about the specific semantic field of each sentence, part of the words were excluded from the sentences. Words such as articles, pronouns, prepositions and auxiliary verbs were excluded while nouns, adjectives and adverbs were kept. Non-words (the sequences of vowel sounds) were also excluded. In Figure 28, the remaining words are displayed in a two-dimensional space. The position of the words in this space was obtained, for visualization purposes only, by reducing the dimension of the word vectors so that similar vectors are close in the reduced space. It shows us that words with related meanings seem indeed are similar in the vector space. For example, the words "France", "pays", "monde" (France, country, world) are grouped together, as well as the words "prix", "chère", "paye" (price, expensive, pay).

The average of the vectors of the words composing a text has proven to be a useful feature for a variety of tasks (Kenter et al., 2016; Farouk, 2019). A common method to evaluate the similarity between average word vectors is the cosine similarity (Kenter and De Rijke, 2015; Gershman and Tenenbaum, 2015). Therefore, a vector representation of each sentence pronounced by P3 was obtained by averaging the vectors associated to its words. Finally, the semantic distance $D_{semantic}$ between two epochs e_i, e_j was measured by the cosine distance between the vectors V_i, V_j of the corresponding sentences:

$$D_{semantic}(e_i, e_j) = 1 - \frac{V_i \cdot V_j}{\|V_i\|\|V_j\|}$$

6.2.7.3 Possible confounders

We decided to take into account 3 additional variables that might influence both the spike train distance and the sentence similarity metrics. The first one is the difference of duration of the two sentences. For a pair of epochs e_i , e_j of durations d_i, d_j, we therefore computed a duration distance $D_{duration}$ given by:

$$D_{duration}(e_i, e_j) = |d_i - d_j|$$

The second possible confounder that was considered is the interval of time separating two epochs during the experiment. Indeed, if the neural activity was slowly modulated over time, the epochs that were recorded around the same time might share similarities. Since repetitions of the same sentences often occurred successively, the associated activity could be expected to show time-related similarities. For the epochs e_i, e_j, starting at time t_i, t_j, the time distance D_{time} is defined as:

$$D_{time}(e_i, e_j) = |t_i - t_j|$$

Finally, the difference of phoneme number was also included as a possible confounder. The phonemic distance was expected to be correlated with this variable. Adding this

[1] http://vectors.nlpl.eu/repository/

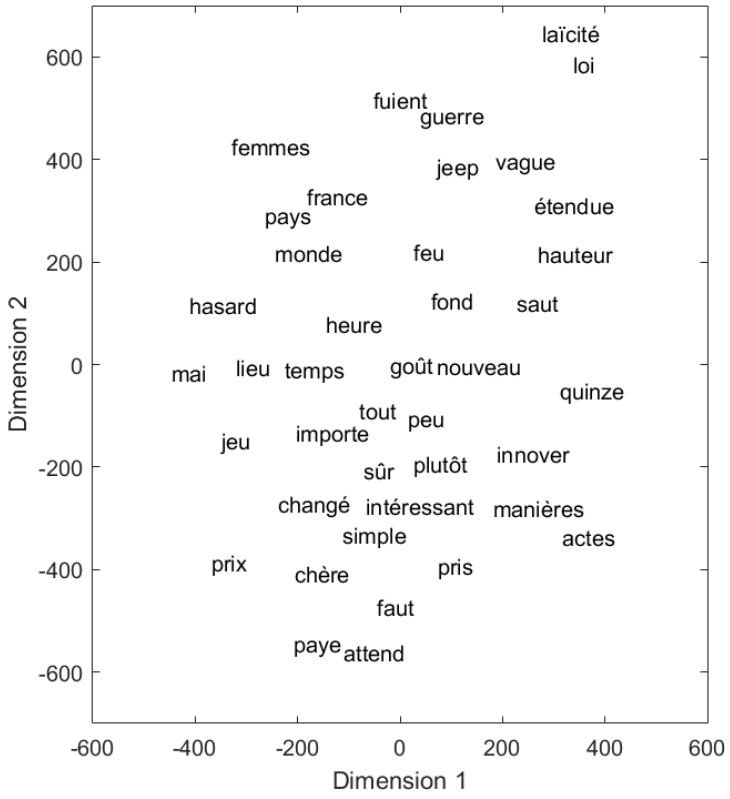


Figure 28. Visualization in 2D of the 100-dimensional vector space of word representations for the considered vocabulary in P3 experiment. The projection in 2D is realized using t-SNE (t-Distributed Stochastic Neighbor Embedding; Van Der Maaten and Hinton, 2008) with the cosine distance as similarity metric and a perplexity parameter of 10. t-SNE is a nonlinear dimension reduction technique that is used to visualize high-dimensional points in 2D or 3D. This algorithm is designed so that similar vectors in the original space are close in the reduced space.

variable in the analysis allowed to evaluate if the additional information captured by the phonemic distance (the nature of the phonemes and their order) was relevant to explain the distributions of spike train distances. For two epochs e_i, e_j corresponding to sentences composed of p_i, p_j phonemes, the phoneme number distance was computed as follow:

$$D_{length}(e_i, e_j) = |p_i - p_j|$$

6.2.8 Spatial mapping

The mapping of the LFP over the Utah array was performed using the NeuroMap software[1] (Abdoun, 2011). This software spatially interpolates the LFP at each time frame using

[1]Freely available from https://www.neurotech-lab.fr/.

thin plate splines.

6.2.9 Relationship between LFP phase and firing rates

The firing rate of each putative unit was examined in relation with the phase of the slow LFP recorded on the same channel. The firing rate and LFP phase were computed on 50 ms bins. The firing rate was computed by counting the number of spikes in each temporal bin and dividing the result by 50 ms. This method was preferred over the use of previously-computed smoothed firing rates because we thought it would allow to evaluate more precisely the presence of a phase-locking phenomenon. To obtain the instantaneous phase, first the analytic signal corresponding to the slow LFP was computed using Hilbert transform. The analytic signal was then averaged for each bin. Finally, the phases of the resulting complex values were computed.

6.3 Results

6.3.1 Speech-related modulation of firing rates

6.3.1.1 Single unit and population modulation

We first compared the average firing rate of each unit during the different speech-related conditions of the participant: overt speech, silence and covert speech. As shown in Figure 29, we found that 9 units over 64 had statistically different firing rates between at least two of the conditions. On a much larger spatial scale, it has been shown that the activity in Broca's area increases during speech compared to resting state and that an increase, although smaller, was also observed during covert speech (see section 3.3). In the present recording, at the single-unit level, such hierarchy of the firing rates between the 3 conditions was not observed.

In order to visualize the modulation of the firing rates along the course of a trial, the firing rates of the 9 modulated units were projected in a 3-dimensional space using PCA and averaged across trials. As can be observed in Figure 30, the average trajectory of the projected firing rates is organized depending on the speech-related condition of the participant during the trials. In other words, the projected firing rates corresponding to overt speech, silence and covert speech tend to be clustered in different parts of the projection space. Statistical testing indicates that the projected firing rates have different overall means for the 3 conditions (one-way multivariate ANOVA, $p < 0.001$; post-hoc pairwise one-way multivariate ANOVA, Bonferroni-adjusted $p < 0.001$). This descriptive analysis of the firing rates shows that the activity of the recorded units seem changed according to the speech-related condition of the participant. In order to reinforce this hypothesis, the firing rates were then classified among the 3 conditions.

6.3.1.2 Classification

The smoothed firing rates values sampled at 20 Hz were classified according to 3 conditions: overt speech, covert speech and silence. The considered features were temporal

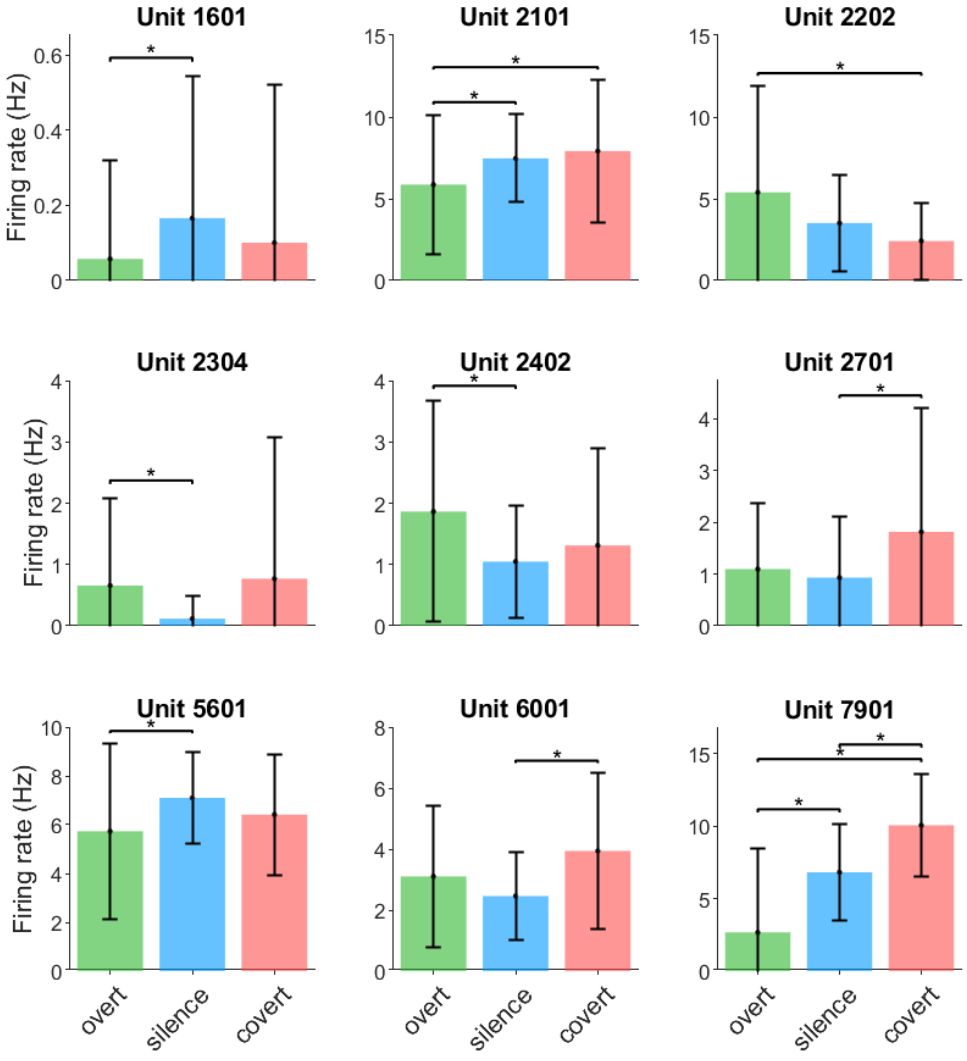


Figure 29. Condition-modulated units. Units modulated by the participant's condition: overt speech, silence or covert speech. The average firing rates were computed in the labeled epochs (defined in section 6.2.5.1). Units which produced different firing rates depending on the condition were identified using Kruskal–Wallis one-way analysis of variance ($p < 0.05$). Post-hoc tests were then computed to assess the modulation between each pair of conditions (Wilcoxon-Mann-Whitney test, Bonferroni-adjusted $p < 0.05$).

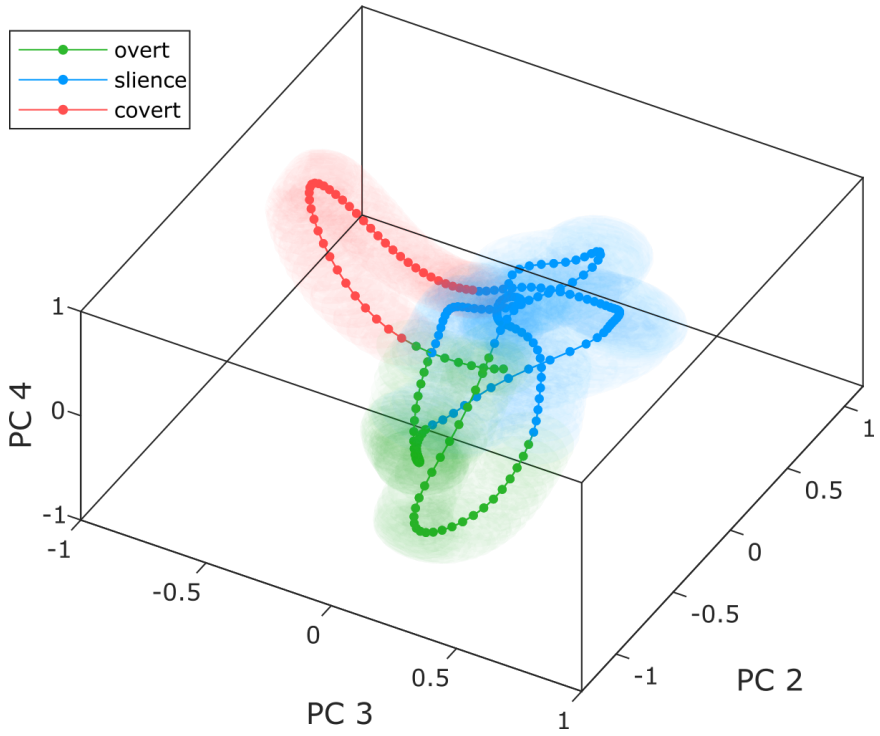


Figure 30. Trial-averaged projection of the firing rates of condition-modulated units using PCA. The firing rates of the 9 modulated units were projected in a 3-dimensional space using PCA. The components 2, 3, 4 were considered as they produced a nicely visible separation of the conditions. The projected trajectories were averaged while preserving the different phases of the trial (see section 6.2.2). The shaded areas are semi-transparent spheres whose radius is equal to the standard error of the mean (SEM). SEM was chosen instead of standard deviation in order to improve readability.

windows of firing rates. LDA classifiers were trained and evaluated in leave-one-out cross-validation based on trials. Feature selection was carried out in each fold according to a statistical evaluation of their discriminative power.

The confusion matrix (Figure 31a) shows that overt and silence samples were correctly classified with an accuracy of approximately 70% (70% and 71% respectively), while the accuracy for covert samples was 53%. The accuracy obtained for the 3 classes are above the theoretical chance level of 33%. The highest ratio of confusions concerned the covert samples misclassified as silence. The evaluation of the classification using Mathews correlation coefficient (MCC) confirmed that the obtained results are way above chance levels (Figure 31b). For each class, the MCC was computed over all samples by considering the binary classification problem of discriminating the chosen class from the two others.

Using the LDA classifiers trained on the labeled data only, the posterior probabilities were computed on the whole trials. These probabilities were then averaged across all trials, as displayed in Figure 31c. These probabilities show indirectly the average variations of the classification accuracy along the course of a trial. On average, the transitions from speech to silences appear to be well predicted by the classifier. Conversely, the output

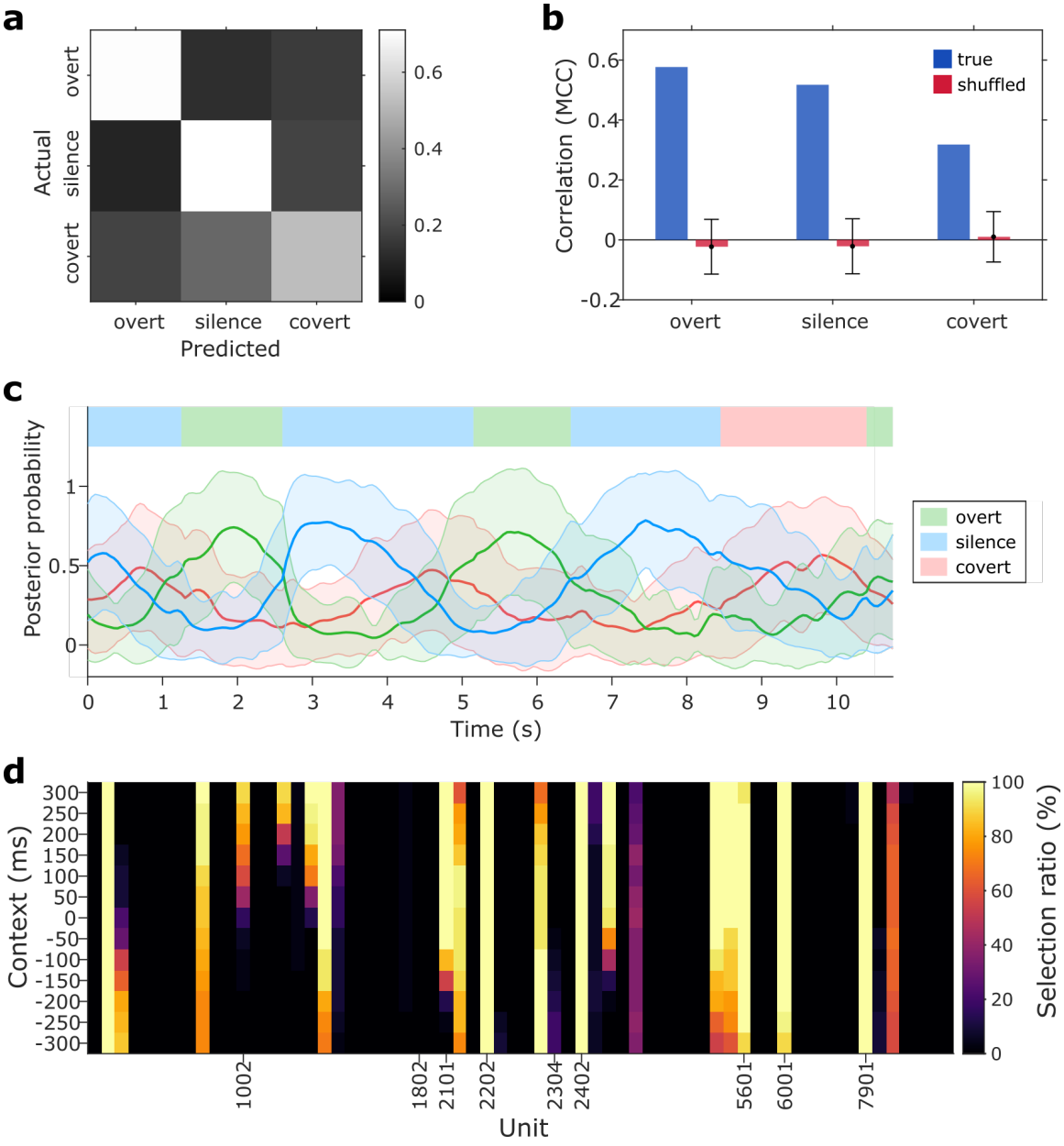


Figure 31. Speech condition classification results. **(a)** Confusion matrix of classification result. **(b)** Evaluation of the performance of the classification for each of the 3 classes using Mathews correlation coefficient. The chance levels, obtained by shuffling the labels of the epochs, were computed 20 times. **(c)** Posterior probabilities of LDA classifier, averaged over all trials. Before computing the average, the posterior probabilities were computed on the whole trials using the classifiers trained on the labeled data. The solid lines and shaded areas indicate the mean and standard deviation of the posterior probability for each class. The colored patches at the top indicate the average phases of overt speech, silence and covert speech. **(d)** Average feature selection during the classification process. The selection ratio indicates the proportion of folds in which a given feature was selected. Positive context refer to the time following the time sample to be classified.

probabilities show that the prediction are much less accurate for the transition from silence to speech. In particular, the samples preceding read and repeat phases by 250 to 500 ms were on average labeled as covert speech instead of silence. This effect could be due to a similarity of the firing rates during covert speech and during the preparation of overt speech.

In each fold, the most relevant features were selected based on statistical tests. The results of this selection, displayed in Figure 31d, show that only a subset of the 64 units were used for the classification. 13 units present features that were used in all the folds, while the features corresponding to 43 units were never used. This average selection on smoothed firing rates only partially overlaps with the selection based on epochs mean firing rates shown in Figure 29. Features corresponding to the activity following the time sample to be classified (positive context) were preferred to preceding activity for a part of the units. The opposite case is observed in less units.

6.3.1.3 Rate coding of phonemes

We evaluated the possibility that the firing rates of some of the recorded units were tuned to certain phonemes, as was observed for neurons of the rostral anterior cingulate cortex (ACC) and the medial orbitofrontal cortex (OFC) by Tankus et al. (2012). Based on automatic segmentation of the audio data, the onset of each phoneme utterance was extracted. Phonemes that had been pronounced less than 20 times were excluded from the analysis, leaving the 13 phonemes /a, ɑ, e, ɛ, i, l, n, o, p, ʁ, s, t, ə/. For each unit and each phoneme utterance, the average firing rate was computed in 3 different time windows relative to the phoneme onset. The time windows covered the time intervals [-300 ms; −100 ms], [-100 ms; 100 ms] and [100 ms; 300 ms]. Statistical tests were run to identify the units that were significantly modulated according to the uttered phoneme in at least one window (Kruskal-Wallis one-way ANOVA, $p < 0.05$). Then, post-hoc tests were computed to assess the modulation between each pair of phonemes (Wilcoxon-Mann-Whitney test, Bonferroni-Holm-adjusted $p < 0.05$). Significant pairwise comparisons were found for 3 units. However, for all 3 units, the firing rates were significantly different for only one pair of phoneme in only one of the 3 time windows. We concluded that no clear tuning to the different phonemes was observed in the recorded firing rates at the level of individual units.

6.3.2 Spike train similarities during speech

After having exposed how the recorded firing rates are modulated depending on the participant's condition, the present section details results obtained from the analysis of spike trains. The spiking activity occurring during overt speech was extracted based on the timing of utterances (manually determined from audio data). Utterances corresponding to both overt reading and overt repetition were included. The extracted data consists of the spike trains of 64 units recorded during 62 overt speech phases. Among all spike trains of all units, half (49%) did not contain any spikes. The non-empty spike trains contained on average 3.5 spikes over a mean speech duration of 1.4 s. One unit was excluded from the analysis as it had spikes in only one of the overt speech epochs.

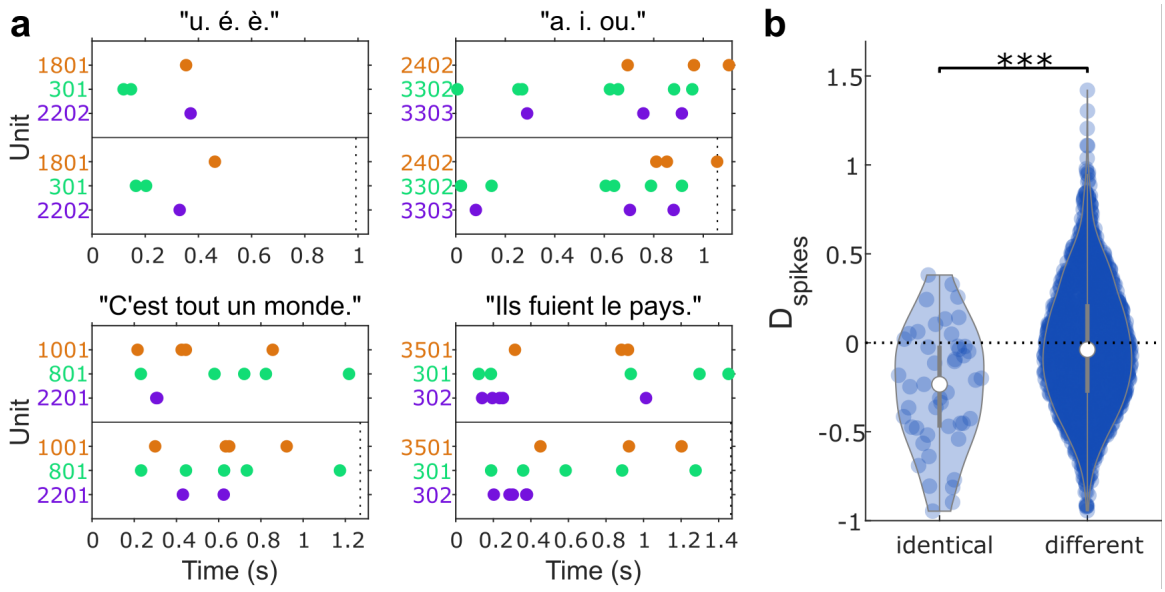


Figure 32. Examples of spike trains and distribution of multi-unit spike train distance for repeated sentences. (a) Example of spike trains occuring while the same sentence is pronounced twice. In each graphs, the raster plots show the activity of 3 units during overt speech epochs. The upper and lower parts of the graphs correspond to two utterances of the same sentence (displayed above the graphs). For each sentence, the activity of the 3 units with the lowest spike train distances are displayed. The x-axis represents the time from speech onset to the end of the longer utterance. The dashed lines indicate the end of the shorter utterance. **(b)** Distributions of multi-unit spike train distances D_{spikes} for pairs of epochs corresponding to identical sentences on the one hand and to different sentences on the other hand. The mean value for identical sentences is statistically lower than for different sentences (one-tailed two-sample Student's *t*-test, $p < 10^{-3}$).

Figure 32 shows examples of spike trains of set of 3 units that occurred during two utterances of the same sentence. The spike trains produced by the selected units during the utterances of the same sentence show similarity in terms of number and timing of spikes. The VP distance was chosen to study the similarity between spike trains as it takes into account both of these aspects of rate and temporal coding. The examples also illustrate the fact that the units that produce the most similar spike trains with respect to the VP distance were different from one sentence to another. We therefore decided to study the similarity of spike trains during overt speech at the population level. The average of the normalized distances of all individual units, noted D_{spikes}, was used as a multi-unit spike train distance.

6.3.2.1 Classification of sentences based on spike train distances

We investigated the possibility to classify spike trains according to the simultaneously uttered sentences using D_{spikes}. Figure 32 shows the distribution of D_{spikes} for pairs of

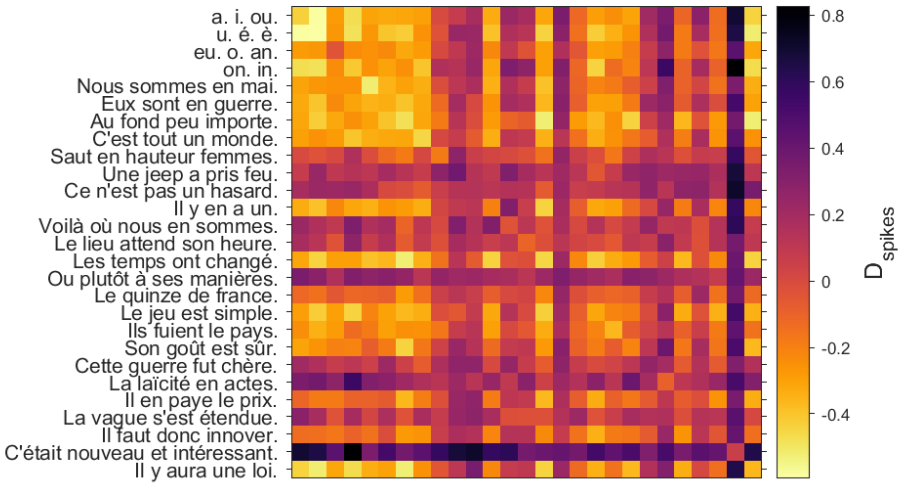


Figure 33. Matrix of average multi-unit spike train distances for the different pairs of sentences. Each value of the matrix is the average multi-unit distance between the spike trains that occurred during the different utterances of two sentences. The diagonal values concern the comparison of spike trains occurring during the utterances of the same sentence. All the sentences except vowel sequences were pronounced only twice in the dataset. For all the sentences except vowel sequences, diagonal values therefore represent the unique distance between the spike trains of the two utterances.

spike trains that were associated to identical sentences on the one hand and different sentences on the other hand. Despite the large overlap between the two distributions, the mean spike train distance for identical sentences was significantly lower than for different sentences (Student's t-test, $p < 10^{-3}$). This result suggests that spike trains recorded during two utterances of the same sentence present similar patterns.

In Figure 33, the spike train distances are summarized in a distance matrix showing the average value of D_{spikes} obtained for the different pairs of sentences. The distance between spike trains corresponding to the utterances of the same sentence are visible on the diagonal. For 8 of the 27 sentences, the average D_{spikes} value was lower for identical than for different sentences, which suggests that D_{spikes} might allow to distinguish some sentences from the others based on the recorded spike trains.

This ability to discriminate the 27 uttered sentences based on the simultaneously recorded spiking activity was ultimately quantified using a simple leave-one-out classification scheme based on D_{spikes}. In each fold, one multi-unit spike train was classified among the 27 sentences based on its distance to the remaining spike trains (see section 6.2.6.2 for more details). As will be discussed in the next section, D_{spikes} is correlated to the difference of duration of the compared spike trains. Since utterances of the same sentence tend to have the same duration, it was preferable to factor out this parameter in order not to bias the classification. Instead of using the end of the utterance, spike trains were therefore extracted from the same 2-second period following the speech onset

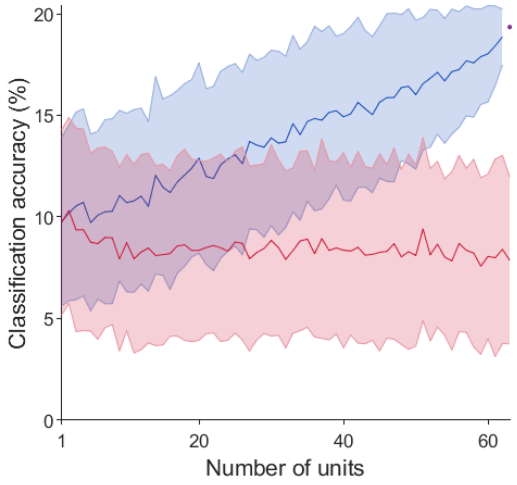


Figure 34. Classification accuracy for different number of units. The curves represent the average accuracy for the classification of sentences based on spike trains using different numbers of units (blue) and corresponding chance level (red). The shaded areas indicate the standard deviation. For each number of units, the classification process was repeated 200 times with randomly selected units. For each random unit selection, D_{spikes} was computed in the same way as previously. The chance level was estimated by shuffling the class labels.

of each utterance. 19.4% of the 62 multi-unit spike trains were correctly classified. This accuracy was found to be significantly over chance level (average chance accuracy of 8%, significance criterion < 0.05).

In order to assess the effect of the number of units on the classification accuracy, the classification process was repeated several times with different numbers of randomly selected units. As can be seen in Figure 34, increasing the classification accuracy is positively correlated to the number of units included in the process. On the considered range, the two variables appear to be linearly correlated ($r = 0.55$, $p < 10^{-3}$).

6.3.2.2 Relationship between spike train distances and sentence similarity metrics

D_{spikes} was computed for the spike trains corresponding to all pairs of utterances. We intended to test the hypothesis that the spike train distance could be linearly related to the similarity of the phonemic and/or semantic content of the compared sentences. To this end, phonemic and semantic and semantic distances were computed for all pairs of uttered sentences. The phonemic distance $D_{phonemic}$, derived from the Damereau-Levenshtein string-edit distance, was designed to take into account both the phonemes composing the sentences and their order. For the semantic distance $D_{semantic}$, a distance based on word vectors, commonly used in computer-based semantic analyses, was chosen to quantify the amount of semantic content shared by pairs of sentences. In the following, the relationships between D_{spikes} and the sentence similarity metrics were examined

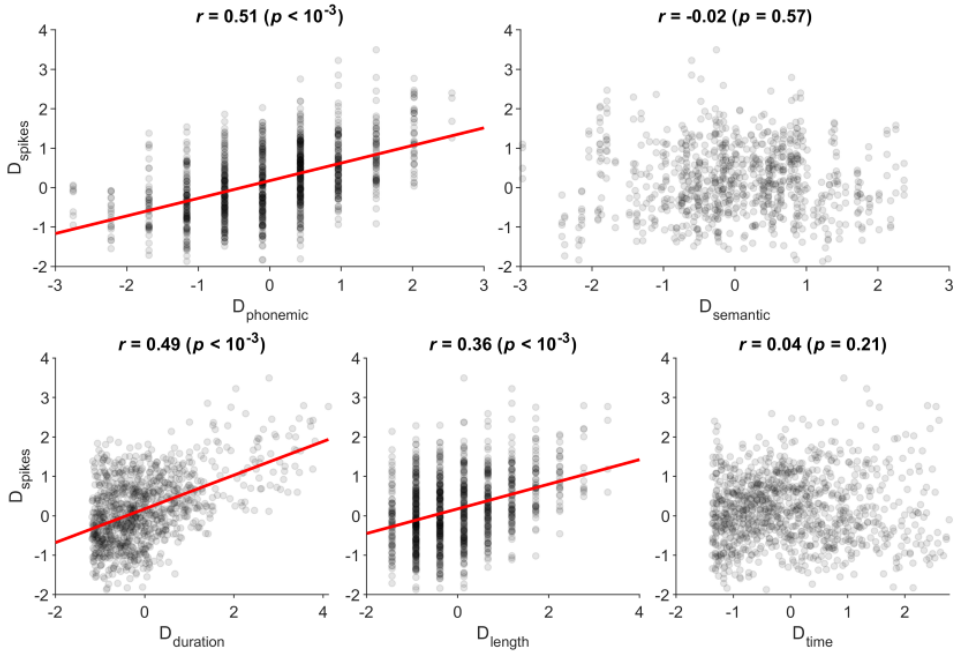


Figure 35. Relationships between the multi-unit spike train distance and the sentence similarity metrics. All variables were normalized to zero-mean and unit-variance. All pairs of epochs associated to different sentences are represented, except for the ones involving a succession of vowels. For all graphs, the Pearson correlation coefficient r and the associated p-value are displayed. The least-square line is displayed in red on the graphs showing a significant linear relationship.

considering only the distances corresponding to the utterance of different sentences. The vowel sequences were excluded of the considered sentences. As shown on the upper graphs of Figure 35, D_{spikes} was found to be significantly correlated with the phonemic distance ($r = 0.51$, $p < 10^{-3}$) but not with the semantic distance.

In order to support the hypothesis that the observed linear relation between $D_{phonemic}$ and D_{spikes} was not due to confounding factors, two additional variables were considered in the analysis: the absolute difference of duration between two utterances, noted $D_{duration}$, and the absolute difference in phoneme number between two sentences, noted D_{length}. Both of them were thought to be likely positively correlated with $D_{phonemic}$ and D_{spikes}. The impact on D_{spikes} of the amount of time separating the compared spike trains, noted D_{time}, was also examined. Our motivation was to rule out the possibility that the variance of D_{spikes} was explained by slow temporal variations of the firing activity over the time course of the recording. As shown on the lower graphs of 35, D_{spikes} was found to be significantly correlated with $D_{duration}$ ($r = 0.49$, $p < 10^{-3}$) and D_{length} ($r = 0.36$, $p < 10^{-3}$), but not with D_{time}.

In order to further verify that the relationship between the phonemic duration and the spike train distances was not entirely explained by the confounding variables $D_{duration}$ and D_{length}, two linear models were fit to the data. First, a simple linear regression was

calculated to predict D_{spikes} based on $D_{phonemic}$. A statistically significant regression equation was found ($F(1, 1010) = 364$, $p < 10^{-3}$), with an R^2 of 0.27:

$$D_{spikes} = 0.45 \cdot D_{phonemic} + 0.16$$

Secondly, a multiple linear regression was calculated to predict D_{spikes} based on $D_{phonemic}$, $D_{duration}$ and D_{length}. A statistically significant regression equation was found ($F(3, 1008) = 186$, $p < 10^{-3}$), with an R^2 of 0.36:

$$D_{spikes} = 0.31 \cdot D_{phonemic} + 0.26 \cdot D_{duration} + 0.07 \cdot D_{length} + 0.16$$

The 3 variables were found to be significantly contributing to the prediction of D_{spikes} ($p < 10^{-3}$ for $D_{phonemic}$ and $D_{duration}$; $p < 10^{-2}$ for D_{length}). The comparison of the two models shows that the contribution of $D_{phonemic}$ was reduced by the inclusion of $D_{duration}$ and D_{length}. It can therefore be inferred that part of the correlation observed between $D_{phonemic}$ and D_{spikes} was due to these confounders. However, $D_{phonemic}$ remained the most contributing predictor, which confirmed that the phonemic content of the sentences was correlated with the observed spike train distances, independently of the duration of the utterances and the phoneme number of the sentences.

6.3.2.3 Spike train similarities during covert speech

D_{spikes} was computed for pairs of spike trains occurring during the covert speech phases. Similarly to overt speech data, the values of D_{spikes} for pairs of different sentences only (with the exclusion of vowel sequences) were analyzed with respect to the sentence similarity metrics. D_{spikes} was found to be individually correlated with $D_{phonemic}$, $D_{duration}$, D_{length} and D_{time}. A multiple linear regression was calculated to predict D_{spikes} based on $D_{phonemic}$, $D_{duration}$, D_{length} and D_{time}. A statistically significant equation was found ($F(4, 248) = 47$, $p < 10^{-3}$), with $R^2 = 0.43$:

$$D_{spikes} = 0.07 \cdot D_{phonemic} + 0.64 \cdot D_{duration} - 0.02 \cdot D_{length} + 0.06 \cdot D_{time} + 0.14$$

Only the coefficient associated to $D_{duration}$ was found to be significantly non-zero ($p < 10^{-3}$ for $D_{duration}$, $p > 0.05$ for the other predictors). It can therefore be assumed that the correlation that was observed between $D_{phonemic}$ and D_{spikes} was due to the confounding effect of $D_{duration}$ and that no significant relationship could be observed between the spike train distances and the phonemic content of the covertly repeated sentences.

The comparison of the analyses of overt and covert speech data remains so-far limited because of the smaller size of the covert dataset and the uncertainty about the start and end times of covert repetitions. In order to produce a fairer comparison, another overt speech dataset was analyzed. This new dataset consisted only of the spike trains recorded during the 31 overt repetitions. For each trial, instead of starting from the speech onset, the spiking activity was extracted starting from the instruction display. The resulting overt repetition dataset therefore had the same size as the covert one and included a similar speech onset uncertainty. As for the complete overt dataset, the spike train distances for pairs of different sentences were found to be significantly correlated to $D_{phonemic}$,

$D_{duration}$ and D_{length}. A multiple linear regression was calculated to predict D_{spikes} based on $D_{phonemic}$, $D_{duration}$ and D_{length}. A significant equation was found ($F(3, 249) = 71$, $p < 10^{-3}$), with $R^2 = 0.46$:

$$D_{spikes} = 0.38 \cdot D_{phonemic} + 0.34 \cdot D_{duration} + 0.14 \cdot D_{length} + 0.08$$

$D_{phonemic}$, $D_{duration}$ and D_{length} were found to be significant predictors of D_{spikes} ($p < 10^{-3}$ for $D_{phonemic}$ and $D_{duration}$, $p < 0.05$ for D_{length}). As for the previous overt dataset, the estimate associated to $D_{phonemic}$ had the largest value. The analysis of comparable overt speech data suggests that the absence of significant relationship observed between D_{spikes} and $D_{phonemic}$ for covert speech is not due to the dataset's size or to speech onset uncertainty. This fact supports the hypothesis that overt and covert spiking activity are different in nature.

To further test this hypothesis, we computed the spike train similarities between the epochs corresponding to overt and covert repetitions, defined from the instruction display as above. Contrarily to what was found for overt epochs (Figure 32b), no statistically significant difference was found between the spike train similarities corresponding to the comparison of identical sentences on the one hand, and different sentences on the other hand (one-tailed two-sample Student's t-test, $p > 0.05$).

6.3.3 Speech-related modulation of slow LFPs

6.3.3.1 Slow LFPs during overt speech

We also analyzes the slow variations of the LFPs recorded with the Utah array. For part of the electrodes, modulations of the LFP below 1 Hz were consistently observed across trials. Figure 36a shows the trial-averaged slow LFP for channels 49 and 66. We verified that these variations are not an effect of the CAR and found that very similar curves are obtained without using CAR in the pre-processing. The averaged signals show that modulations occur during and after overt speech. On channel 49, a decrease of the potential during speech is followed by an almost symmetric increase after speech. On channel 66, similar variations are observed but with an inverse polarity. These slow modulations related to overt speech could be observed on single epochs, as shown in Figure 36b. It shows that for channels 49 and 66, consistent modulations are observed for most of the overt speech epochs in the dataset. For both channels, the duration of the peaks did not seem to be related to the duration of the overt speech phase.

The average slow LFP modulations during overt speech are visualized over the whole Utah array in Figure 37. It reveals that slow variations of potential were observed in an important number of electrodes. Most of the electrodes display a positive variation during speech but negative variations were observed in several electrodes. On part of the electrodes exhibiting a modulation, an inversely polarized "rebound" seems to be observed subsequently. The average peak time seems to be different depending on the electrodes. On some of the displayed frames, modulations seem to form spatially coherent patterns – in particular the positively modulated electrodes in the center of the array at 1 s after speech onset and the negatively modulated electrodes on the right side of the array at 1.5 s

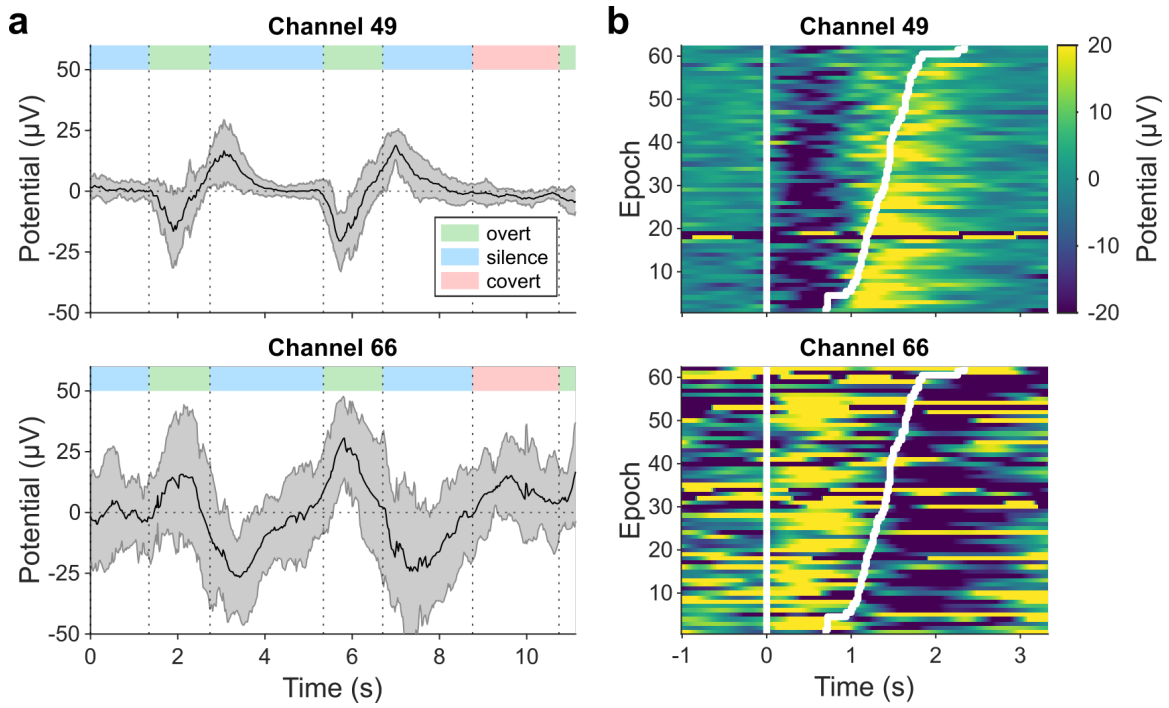


Figure 36. Example of slow LFP variations for channels 49 and 66. (a) Trial-averaged slow LFPs. For each channel, the solid line indicate the median potential and the shaded area represents the MAD-based standard deviation. These statistics were chosen as they are more robust to outliers than mean and standard deviation. The colored patches at the top indicate the average phases of overt speech, silence and covert speech. **(b)** Slow LFPs during overt speech epochs. Each line represents the color-coded amplitude of slow LFP for a given utterance. The epochs are sorted with respect to the duration of the utterance. The first white line indicates speech onset. The second white line indicates the end of each utterance.

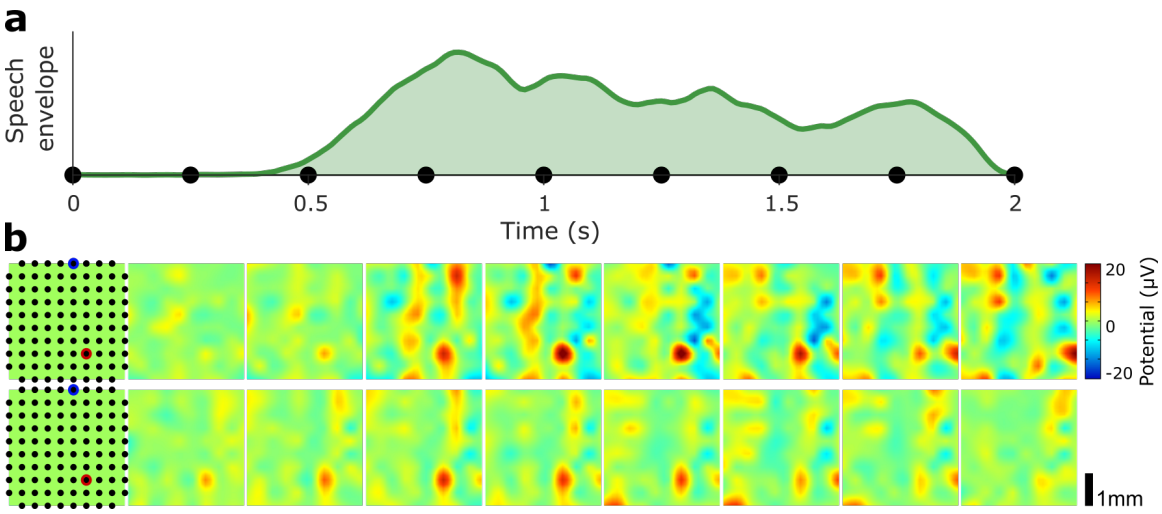


Figure 37. Mapping of the average slow LFP on the Utah array during overt and covert repetitions. Data from both repetitions was extracted between the time of the instruction display and the end of the utterance. The end of the utterance is assimilated as the beginning of the "OK" for the covert condition. The extracted epochs were stretched or compressed to match the average duration of 2 s. **(a)** Average of the speech envelope for overt repetitions. The speech envelopes were computed using Hilbert transform. **(b)** Spatial mappings of the average slow LFP over the Utah array for overt repetitions (upper maps) and covert repetitions (lower maps). The maps represent the amplitude of potential at different times after the instruction display. The maps correspond to the times indicated by black dots in panel a. On the first maps, the position of the 96 electrodes are displayed. Adjacent electrodes are 400 µm apart. The electrodes circled in blue and red represent channels 49 and 66, respectively.

after speech onset. However it overall seems that these modulations are very localized.

6.3.3.2 Slow LFPs during covert speech

As can be seen on Figure 36, channel 66 whose potential increases on average during overt speech, seems to display an attenuated but similar modulation during covert speech. Comparison between overt and covert repetitions (Figure 37b) also shows a similar increase of the potential on this channel, peaking at about the same time after the instruction display. Other electrodes seem to show attenuated positive modulations similar to overt speech. The overall spatial distribution of potential over time seems to be similar in both cases but with lower amplitudes in the covert speech condition. The similarity between the two conditions remains limited at the end of the epochs, possibly because the covert speech was followed by the utterance of the "OK", which probably influences the preceding activity.

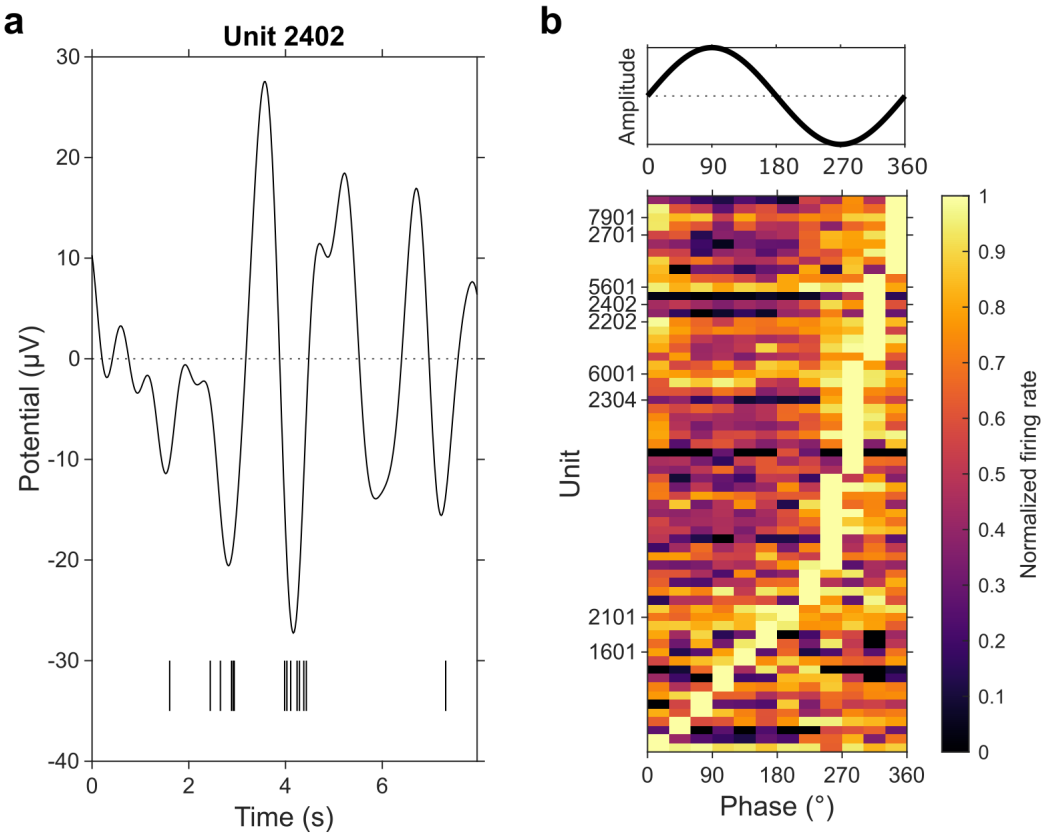


Figure 38. Relationship between firing rates and slow LFP phase. (a) Extracts the LFP variation on channel 24 and the spiking activity of unit 2402 that was recorded on this channel. **(b)** Each line of the colormap represents the mean firing rate of a unit with respect to the phase of the LFP of the corresponding channel. The firing rates were normalized for each unit so that the maximum of the mean firing rates equals one. The units were sorted with respect to the phase of at which their firing rate is maximal. The upper graph illustrates the typical variations of a signal in relation to phase.

6.3.3.3 Relationship between firing rates and slow LFP phase

Figure 38a shows an extract of the slow LFP recorded on channel 24 and the spiking activity of unit 2402, recorded on the same channel. It appears on this extract that the detected spikes tend to occur during the negative peaks of the slow LFP. Figure 38b shows the mean firing rate of each unit with respect to the phase of the LFP on the same channel. As suggested the extract on panel a, unit 2402 fired the most during the negative peaks of the LFP oscillations. It appears that a majority of units (77%) had a maximum firing rate for phases between 180 and 360°, which correspond to the negative peak of the oscillations. These units also tend to show a decrease of the firing rates associated with the positive peak of the oscillations.

6.4 Discussion

In this study, we examined the activity of a neuronal population in Broca's area during speech production tasks. The firing rates of the recorded units were on average inferior to 8 spikes/s which is low compared to the firing rates reaching tens of spikes per second that have for example been reported in the ACC and OFC during speech production (Tankus et al., 2012) or in the dorsal motor cortex during hand movements (Hochberg et al., 2012).

We found that the firing rates of the recorded units were differently modulated during overt and covert speech. It was possible to differentiate silence, overt and covert speech using a classifier based on the smoothed firing rates. Most of the confusions were observed primarily between covert and silence, and then between overt and covert (Figure 31a). Contrarily to several ECoG studies, which measured peaks of activity in Broca's area over a 500-ms period before speech onset (Flinker et al., 2015; Brumberg et al., 2016), we found that, for most of the modulated units, future activity was more informative to classify the cognitive state at a given instant (Figure 31d), implying that they are modulated after speech onset. Finally, the output obtained by applying the classifier to whole trials revealed possible similarities between the neural activities during covert speech and preceding overt speech onset.

Using a multi-unit spike train distance that took into account both rate and temporal coding, it was possible to classify spike trains according to the simultaneously pronounced sentence over chance level. However, following the trend visible in Figure 34, hundreds of units would be required to classify the 27 sentences with a satisfying accuracy. This conjecture should be nuanced by the fact that the choice of classifying algorithm was limited by the low number of instances of each class. Additional recordings, featuring more covert repetitions of the same sentences, would be necessary to investigate the possibility of classifying covertly produced sentences based on neuronal activity in Broca's area.

While no clear tuning of the firing rates to specific phonemes was observed, analyses revealed that the multi-unit spike train distance was correlated to the phonemic content of the pronounced sentences. This could mean that the phoneme information is indeed encoded in the recorded activity, but not in the firing rate of individual neurons. Additional experiments, focusing on the production of phonemes, would be necessary to investigate the type of coding that might be involved. No correlation was observed between the multi-unit spike train distance and the semantic content of the pronounced sentences, which is coherent with the claim of most studies that Broca's area deals with lower level speech components. For covert speech segments, no correlation was observed between the multi-unit spike trains distance and the phonemic content of the sentences. No evidence was either found that overt and covert repetitions of the same sentences elicited similar spike trains.

In the same intracortical micro-electrode recording, we have observed modulation of the slow LFP ($< 1\,\text{Hz}$) during overt speech. Slow variations of potential in the cortex have been reported in motor areas in relation to motor tasks under the denomination of local motor potential (LMP) or movement-evoked potential. Such signals, recorded using intracortical micro-electrode or ECoG, have even been used for the control motor BCIs by

non-human primates and human subjects (Flint et al., 2013; Stavisky et al., 2015; Gunduz et al., 2016). However, this is, to our knowledge, the first study that the local modulation of such slow rhythms is reported in Broca's area. We have observed that most of the recorded units tended to fire during the negative peaks of the oscillations. In future works, it would be interesting to investigate if this relation between the spiking activity and slow LFP phase in Broca's area encodes speech features, similarly to phenomenons like place coding in the hippocampus (O'Keefe and Recce, 1993).

Altogether, the results obtained on the smoothed firing rate, on the spike train distances and on the slow LFP support the hypothesis that the neuronal activity of the recorded units is different during overt and covert speech production. We were able to differentiate the spiking patterns of different sentences but a larger number of recorded neurons and a better understanding of how speech features are encoded would be necessary to build a speech BCI based on the neuronal activity of this area. Finally, it should be noted that the main limitation of this study lies in the fact that it is based on the neural recording of a single subject. Additional recordings from different participants would be necessary to know if the studied recording is representative of the activity found in Broca's area. Considering the multiplicity of the roles attributed to Broca's area, it could also be expected that the neuronal activity varies significantly depending on the location within Broca's area.

Chapter 7

Preliminary decoding of covert speech based on electrocorticography

7.1 Introduction

The progresses in continuous decoding of speech invite to envision the possibility of building a real-time speech BCI in the near future. This interface could be achieved by converting features of the neural activity into features of the intended speech. Two different approaches exist to continuously characterize speech over time: acoustic and articulatory representations. In both approaches, compact representations with less than 30 dimensions are sufficient to re-synthesize intelligible speech (Bocquelet et al., 2016b; Chartier et al., 2018; Akbari et al., 2019). Studies decoding continuous articulatory or acoustic representations from brain activity during speech production have obtained promising results (Guenther et al., 2009; Martin et al., 2014; Anumanchipalli et al., 2019). In particular, Anumanchipalli et al. (2019) reported better performances when decoding articulatory speech features and were able to synthesize intelligible sentences from the decoded articulatory trajectories.

In most of the existing studies, speech is decoded from the neural activity of subjects during overt speech production (see Table 2). However, additional challenges must be faced in order to decode the intended speech of people that are unable to articulate due to paralysis. In particular, the activity patterns recorded in paralyzed participants while they intend to speak would probably differ from the activity of healthy participant during actual speech production. Indeed, the somatosensory feedback present in healthy subjects while they speak contribute to the recorded brain activity. Regarding the absence of somatosensory feedback, the neural activity elicited by covert speech in healthy participants is a relevant model of what might be observed in paralyzed subjects during attempted speech production. As discussed in section 3.3, the neural activity patterns observed during overt and covert speech share some similarities but lower activations have been observed in the auditory and motor areas (Palmer et al., 2001; Pei et al., 2011b). In Chapter 6, we showed that the spiking activity of a population of neurons was different during overt and covert speech. In most decoding studies comparing the covert (or mimed) speech with overt speech, it also appears that imagined or mimed speech is decoded with a lower

accuracy than audible speech (Martin et al., 2014, 2016; Anumanchipalli et al., 2019).

Another obstacle to the decoding of covert speech comes from the absence of behavioral output. Indeed, most of the speech decoding algorithm are based on corpora of neural activity temporally aligned with the speech features to decode. Martin et al. (2014) and Anumanchipalli et al. (2019) studied the possibility to decode covert and mimed speech by training the decoder on overt speech data. However, such calibration would not be possible with participants unable to speak. An alternative strategy was used by Guenther et al. (2009) to decode speech formants based on the neuronal activity of a participant suffering from locked-in syndrome. In their study, the decoder was calibrated using data acquired while the participant was asked to imagine pronouncing vowel sounds in synchrony with an audio stimulus.

In this chapter, preliminary results concerning the continuous decoding of speech features during covert phoneme production are presented. These results are based on the electrophysiological recording of one human participant, who imagined pronouncing individual phonemes continuously for several seconds. This protocol allowed to calibrate of the decoder on covert data. First, the results of the offline decoding of articulatory and acoustic features of the imagined phonemes are presented. We show that the acoustic representation was decoded over chance level using very low frequency components of the LFP in Broca's area and the superior temporal gyrus. In a second part, a first attempt to decode covert speech production in real-time via the articulatory representation is described.

7.2 Material and methods

7.2.1 Subject

This study is based on part of the data from participant P5 (described in details in section 5.2). This participant was implanted with a 72-electrode grid covering the left hemisphere (Figure 39) for several days as part of a presurgical evaluation of her intractable epilepsy. She was also implanted with a 96-electrode Utah array in the left ventral sensorimotor cortex. She took part in several recordings following different protocols. This study is based on one closed-loop decoding session and one open-loop recording sessions during which she was instructed to perform covert production of phonemes.

7.2.2 Set of phonemes

The phonemes used in this study were a subset of 21 French phonemes selected based on the fact that they can be continuously pronounced with a sustained position of the articulators. The set of phonemes was composed of oral vowels /a, i, u, y, e, ɛ, ə, o/, nasal vowels /ɑ̃, ɔ̃, ɛ̃/, voiced consonants /m, n, v, z, ʒ, l, ʁ/ and unvoiced consonants /f, s, ʃ/.

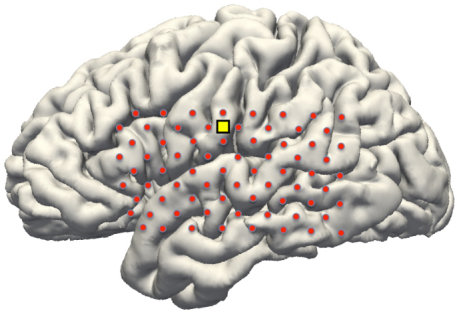


Figure 39. Localization of the ECoG electrodes and Utah array for participant P5. The ECoG electrodes (red dots) and Utah array (yellow square) are displayed on the brain reconstruction obtained from pre-operative MRI.

7.2.3 BY2014 articulatory-acoustic corpus

The BY2014 corpus is an articulatory-acoustic corpus previously acquired by Blaise Yvert's team (Bocquelet et al., 2016c) and made freely available[1]. It contains audio signals and articulatory trajectories of a male French speaker reading more than 650 items. The items include isolated vowels, vowel-consonant-vowel sequences (VCVs), phonetically balanced sentences, and other sentences extracted from articles of the French newspaper *Le Monde*. The positions over time of the main articulators were recorded using EMA (see section 2.3.2). The extracted articulatory trajectories consist in the positions in the sagittal plane of 7 landmarks: the back, middle and tip of the tongue, the jaw, the upper and lower lips, and the velum.

7.2.4 Acoustic and articulatory representations of phonemes

For each of the 21 phonemes considered in the study, typical acoustic and articulatory representations were extracted from the BY2014 corpus. The 14 articulatory parameters are the positions in the sagittal plane of 7 relevant points of the vocal tract recorded using EMA (see section 2.1.2.3).

Acoustic parameters were computed from the audio signals of the BY2014 corpus in the form of mel-cepstral filter coefficients (MELs; see section 2.1.3.2). Specifically, 25 MELs coefficients were extracted from 50-ms windows separated by 10-ms time intervals using the SPTK toolkit (SPTK Working Group, 2017).

The sentences of the BY2014 corpus were segmented into phonemes based on the audio recording. This process used hidden Markov models on MFCCs (seen section 2.1.3.2) and was implemented using the HTK toolkit[2]. The segmentation was used to extract the trajectories of the acoustic and articulatory parameters corresponding to the all the instances of the 21 considered phonemes. For each phoneme, the mean acoustic and articulatory trajectories were computed by averaging all instances. In order to do so,

[1] https://doi.org/10.5281/zenodo.154083
[2] Available at https://htk.eng.cam.ac.uk/.

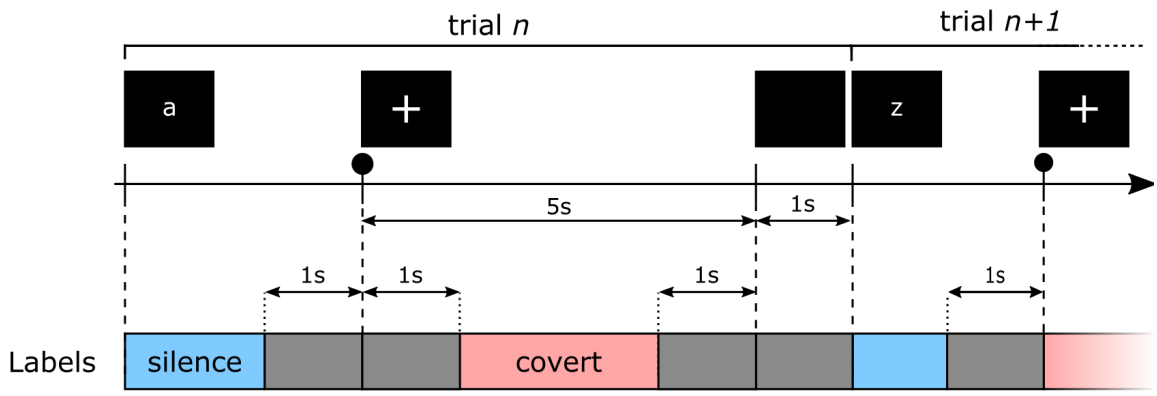


Figure 40. Course of a trial for the covert phoneme production experiment performed by participant P5. The black squares represent the visual cues that were displayed on a screen to the participant. The timeline indicate the time of appearance of the different visual cues. The events marked with a black dot were triggered by the participant pushing a button. The colored boxes show how the trials were divided into different labeled interval for the classification. The grey boxes indicate intervals that were excluded from the classification.

all instances had been previously made of equal duration using linear interpolation. For each phoneme, typical acoustic and articulatory parameters were extracted by taking the midpoint value of the mean trajectories. As a result, each of the 21 phonemes p was associated with speech representation vectors. The value of these vectors for parameter k is denoted $R_p(k)$, where k refers to one of the 25 coefficients in the case of acoustic representation or one of the 14 positions in the case of articulatory representation.

7.2.5 Offline open-loop decoding

7.2.5.1 Protocol

In each trial of this protocol, the participant was asked to imagine pronouncing individual phoneme displayed on a screen. The experiment was divided into 3 blocks of 21 trials. In each block, the 21 phonemes described in section 7.2.2 were presented in a random order. The participant was instructed to push a button when she was ready to covertly pronounce the displayed phoneme. Pushing the button triggered the display of a fixation cross for 5 s. The participant was instructed to imagine pronouncing the phoneme until the fixation cross disappeared, without making actual articulatory movements. Audio and video were recorded to verify the absence of audible sound and visible articulatory movement during the display of the fixation cross. After the cross had disappeared, the screen stayed blank for 1 s before the start of the next trial.

7.2.5.2 Task labeling

In order to classify neural activity with respect to covertly produced phonemes, the data of the different trials was labeled. The epochs corresponding to the display of the fixation cross, with the exclusion of the first and last 1-s intervals, were labeled according to the phoneme (see Figure 40). The first 1 s intervals were not considered because they might contain activity related to the button push. The last 1-s intervals were not considered because the participant might have anticipated the end of the task.

The epochs corresponding to the display of the phoneme, with the exclusion of the last 1 s intervals, were labeled as silence. The last 1 s intervals were not considered because they might contain activity related to the button push.

7.2.5.3 Neural data processing

The signals of the 72-electrode ECoG grid were recorded together with the audio from the microphone at 30 kHz. The signal of one noisy channel was not included in the analyses. Median CAR was applied to mitigate the effect of external artifacts, in particular powerline noise. The signals were downsampled at 2 kHz (after using an anti-aliasing low-pass filter with a 500-Hz cutoff frequency).

A time-frequency representation of the signal of each channel was obtained using continuous wavelet transform (CWT) using generalized Morse wavelets. The obtained power spectrogram was defined for 81 pseudo-frequencies ranging from 0.2 to 199 Hz.

The spectrogram of each channel was z-scored using the mean and standard deviation computed over the silence-labeled intervals. The z-scored power spectrograms was averaged in 9 frequency bands: 0-1, 1-5, 5-10, 10-15, 15-20, 20-30, 30-40, 40-90, 90-200 Hz. The resulting neural features had 71 channels $N = \times$ 9 bands $=$ 639 dimensions. Finally, these band powers were smoothed over time by computing the moving average on a 200-ms window. No features were added for temporal context.

No clear spike waveforms could be extracted from the signals recorded with the Utah array. As a consequence, these signals were not considered in the offline analysis.

7.2.5.4 Estimation of acoustic and articulatory representations of the imagined phonemes

In the neural recording, the intervals of covert production were associated with typical acoustic and articulatory representations corresponding to the imagined phoneme. The process to obtain these typical representations R_p for each phoneme p is described in section 7.2.4. As a result, each sample of neural features (within the covert speech production intervals) was associated to estimated speech features. The estimated speech representation at time t and parameter k is denoted $R(t, k)$, where k refers to one of the 25 coefficients in the case of acoustic representation or one of the 14 positions in the case of articulatory representation. For all samples t during the covert production of phoneme p, $R(t, k) = R_p(k)$. Each parameter of these speech features R was normalized to zero-mean and unit-variance over the whole dataset. Examples of the resulting estimated MEL coefficients for different trials are visible on the upper panel of Figure 46.

7.2.5.5 Regression

Decoding models, mapping the selected features on the speech representation (acoustic or articulatory), were separately computed using multiple linear regression. Each feature of the speech representation was modeled as linear weighted sums of the neural features:

$$\hat{R}(t,k) = \sum_{n=1}^{N} M(n,k) \cdot F(t,n)$$

In this formula, $\hat{R}(t,k)$ is the decoded speech representation at time t and parameter k. $F(t,n)$ is the neural feature n at time t. n refers to the power of the LFP in one of the frequency band of one of the electrodes ($1 \leq n \leq 639$) . Finally, M is the linear regression matrix. The regression matrices were estimated using QR decomposition.

7.2.5.6 Features selection

In order to limit overfitting of the training set, features selection was carried out in each fold. A linear regression model was first computed using all the neural features. Then the decoding contribution C of each neural feature n was computed as the L^2-norm of the associated regression weights across all speech features:

$$C(n) = \sqrt{\sum_{k} M^2(n,k)}$$

A subset of N_s neural features with the highest contributions were selected and the definitive decoding model was obtained by computing the linear regression using only these features. The different decodings were evaluated for values of the parameter N_s between 10 and 200.

7.2.5.7 Cross-validation

The covert-labeled intervals of all trials were included in the decoding analysis. The decoding performance was evaluated using leave-one-out (LOO) cross-validation on the different trials. In each fold, using the trials of the training set, the neural features were normalized to zero-mean and unit-variance, the N_s most contributing features were selected and the regression matrix was computed. The same normalization parameters, selection and regression were then applied to the neural features of the remaining trial. This process was carried out independently to decode the acoustic and articulatory representations on the whole dataset. The evaluation of the decoding was performed by computing the Pearson's correlation between the original and decoded speech representations on the data of all trials put together.

7.2.5.8 Classification of the decoded acoustic representation

We evaluated the ability to classify the decoded speech representations $\hat{R}$ into the different phonemes. Each phoneme class p was represented by its typical feature vector R_p. The

classifier attributed a decoded phoneme $\hat{p}$ to each sample of $\hat{R}$ based on the closest class, according to the Euclidean distance:

$$\hat{p}(t) = \min_{p} \sqrt{\sum_{k} (\hat{R}(t,k) - R_p(k))^2}$$

Since the different classes were perfectly balanced in the dataset, the classification was simply evaluated using the accuracy.

7.2.5.9 Chance level estimation

The statistical significance of the correlation and classification results was assessed by permutation tests. Chance distributions were estimated by carrying out 100 times the decoding on randomized data, obtained by shuffling the phonemes corresponding to the different trials. The statistical significance of the results was assessed by computing the proportion of chance estimations that were superior to the considered results. This proportion will be referred to as the p-value of the permutation test.

7.2.6 Online closed-loop decoding

7.2.6.1 Protocol

Participant P5 also took part in a closed-loop paradigm attempting to decode the activity underlying covert production of sustained phonemes in real-time. A subset of 11 phonemes were considered in the experiment: vowels (/a, i, u, y, e, ɛ, ə, o/, /ɑ̃, ɔ̃, ɛ̃/) and voiced consonants /m, n, v, z, ʒ, l/. In each trial, a phoneme was randomly selected and the participant was verbally instructed by the experimenter to imagine pronouncing this phoneme. She was asked to start imagine pronouncing the requested phoneme and to continue doing it while pressing on a button (Figure 41). While the button was pressed, her neural activity was decoded in real-time and an auditory feedback was provided. She was asked to try to make the auditory feedback match the target phoneme.

7.2.6.2 Software

The neural activity of the participant was processed in real-time by a program called PulsIO, developed in Blaise Yvert's team by Florent Bocquelet and Marc Aubert. The program, coded in C++, provides a framework to design the real-time processing of input data. The processing chain is represented as a graph consisting of connected blocks, as visible on the screenshot of the graphical user interface (GUI) in Figure 42. The blocks represent sources, processing units or sinks. The sources emit packets of data, the sinks receive those packets and the processing units do both. Sources were used to interface the acquisition systems and retrieve in real-time the neural data as well as the button pressing signal. Processing units performed the operations required to extract neural features, decode a speech representation and synthesize audio. Sinks were used to write the data emitted by different blocks in files and to send the synthesized audio to the computer's sound card.

Figure 41. Photograph of participant P5 during the closed-loop experiment. The photograph shows participant P5 while she imagined pronouncing a phoneme. She was pressing the button with her left thumb to activate the decoding process. Audio feedback was provided by the speaker on her left. A microphone simultaneously recorded the sound.

In PulsIO, each type of block is implemented by a C++ class that fits into the structure of the core software. My work mainly focused on the development of the processing blocks dedicated to the control of the experiment (button input, choice of target phoneme), the spike detection and the decoding model (training and application). The other blocks used in the closed-loop experiment were developed by Florent Bocquelet and Marc Aubert. A simplified version of the graph used in the closed-loop experiment is presented on Figure 43.

7.2.6.3 Features extraction

Both ECoG and intracortical multi-unit activity were used to extract features for the decoder. The signals of the 72 ECoG channels were sampled at 2000 Hz by the acquistion system and retrieved in PulsIO. Common average reference was applied. The spectrogram of each channel was computed every 10 ms using a Fourier transform on 256-ms windows. The power of each frequency bin was normalized to zero-mean and unit-variance using exponentially weighted moving estimations (600-s time constant). Band powers were computed by averaging the spectral power in the following bands: 5-15, 15-30, 30-40, 60-100 and 100-200 Hz.

The signals of the 96 micro-electrodes were sampled at 30 000 Hz by the acquisition system and retrieved in PulsIO. A band-pass finite impulse response (FIR) filter was applied to isolate the frequency content between 400 and 3000 Hz. The standard deviation of each channel was computed (runing estimation with a 60-s time constant). The MAD-based estimation of the standard deviation was preferred for the reasons explained in section

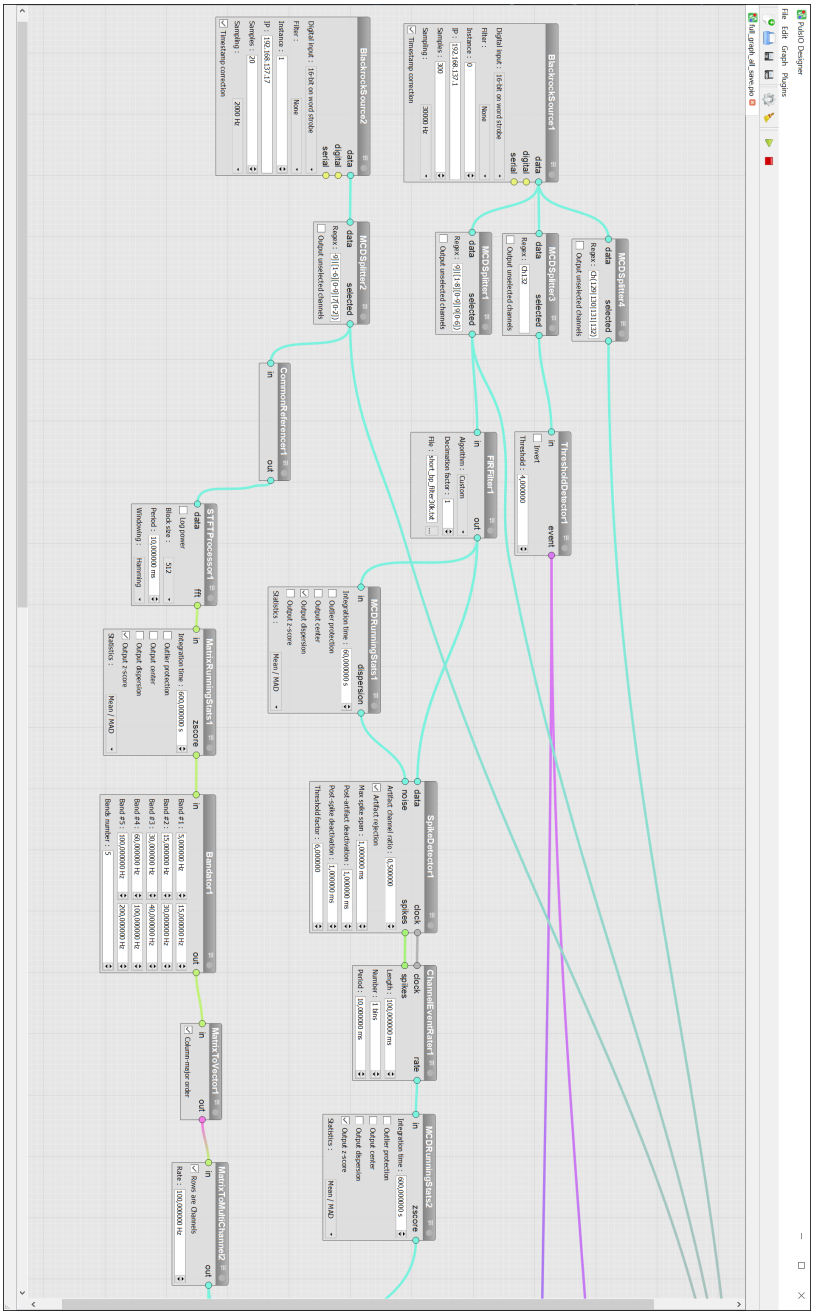


Figure 42. Screenshot of the GUI of the PulsIO software used for closed-loop decoding used for closed-loop decoding. The processing chain is represented as a graph constituted by connected blocks. The blocks emit and receive packets of data. The connections between blocks are represented by colored links whose color depend on the type of transiting data. Links connected to ports on the right side of a block indicate data emissions, while ports on the left indicate reception. The screenshot partially shows the graph that was used in the closed-loop experiment. The two leftmost blocks represent interfaces with the neural data acquistion systems. Most of the blocks are designed to display editable parameters, which can be modified during the experiment. The screenshot shows a part of the graph that was used in the closed-loop experiment.

5.3.1.4. The thresholds for spike detection was set as plus or minus 6 times the standard deviation, based on observation during the recording sessions. Threshold crossing periods lasting more than 1 ms or appearing on more than half of the channels at the same time were considered as artifacts, and otherwise as spikes. After the detection of a spike on a given channel, threshold crossings were not considered during a refractory period of 1 ms. A refractory period of 100 ms was applied after artifact detection. The spiking rate was computed every 10 ms based on the last 100-ms period. The spiking rate of each channel was normalized to zero-mean and unit-variance (running estimation with a 600-s time constant).

The band powers of the ECoG channels and the multi-unit spiking rates of the Utah channels were concatenated to form the neural features.

7.2.6.4 Closed-loop calibration

Using one of the processing block ("Choice of target" in Figure 43), the experimenter could randomly select one of the 11 phonemes. After being instructed about the selected phoneme, the participant could decide when to trigger the real-time decoder using a push button.

Before the experiment, typical articulatory positions had been stored for each phoneme p, in the form of a 14-dimensional vector R_p (see section 7.2.4). While the participant pressed the push button, the samples of neural features were stored in a neural data buffer and the articulatory positions corresponding to the requested phoneme were stored in another buffer. We implemented the buffers so that the data they contain would be balanced according to the different phonemes in case of prolonged use. The neural and articulatory buffers were divided in 11 parts, one for each possible phoneme. Each part was a circular buffer that could store 60 s of data. During the production of a given phoneme, only the corresponding parts were filled. This buffer system was inspired by the study of (Brandman et al., 2017) on closed-loop BCI for 2D cursor control.

Every second, a linear regression model was trained using all the data in the buffers, to predict the articulatory positions based on the neural features. The linear model consisted in a regression matrix obtained by computing 14 ridge regressions, one for each articulatory component to decode. The regularization parameter λ could be fixed or optimized. In the second case, every time that the model was computed, the regularization parameter λ of each ridge regression was optimized. The optimization consisted in selecting the best λ value (among $10^{-8}, 10^{-7}...10^{2}$) according to the L-curve criterion, as presented by Hansen (2000). All regression equations were solved using LU decomposition. Filling the buffers and training the model stopped when the participant released the button, usually after approximately 5-10 s. Then, a new phoneme was selected by the experimenter.

7.2.6.5 Decoding and synthesis

While the button was pressed by the participant, the neural features were continuously decoded using the last computed regression matrix. The estimation of articulatory positions $\hat{R}$ were fed to an articulatory speech synthesizer.

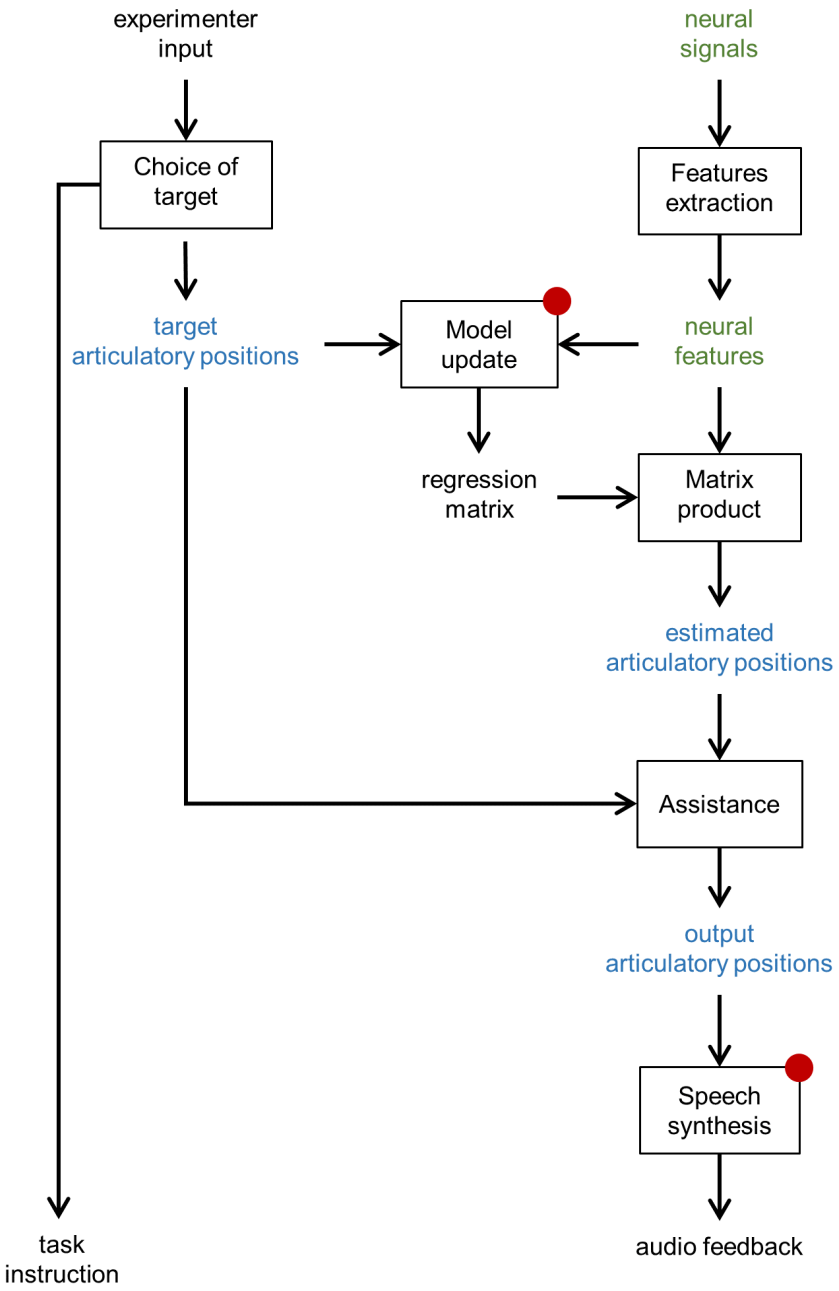


Figure 43. Schematic diagram of the online decoder. The processing blocks marked with a red dot were only active when the participant pressed the button. The "Model update" box contains buffers storing the target articulatory positions and the corresponding neural features.

An extra processing step, referred to as "Assistance" (Figure 43), was added between the decoding and the synthesis. The purpose of this operation was to make the estimation closer to the target positions in order to provide a correct audio feedback at the beginning of the experiment, when the model was being trained on little data. The resulting articulatory positions $\hat{R}_{output}$ were defined as a weighted sum of the target R_p and the estimation $\hat{R}$:

$$\hat{R}_{output} = \alpha R_p + (1 - \alpha)\hat{R}$$

The level of assistance was defined by the parameter α. The value of α was gradually decreased from full assistance ($\alpha = 1$) to no assistance ($\alpha = 0$) along the course of the experiment.

Speech sound waveforms were synthesized from the articulatory positions $\hat{R}_{output}$ in a two-step process. First, an artificial neuron network (ANN) was used to convert the articulatory positions to acoustic coefficients (MELs). The process to build the ANN from the BY2014 corpus is detailed in Bocquelet et al. (2016c). Then, the MELs were used as the coefficient of a digital MLSA filter to generate the waveforms (see section 2.1.3.2).

7.3 Results

7.3.1 Band power modulation during covert speech

Using the data from the open-loop recording, we first observed the average cortical activation pattern corresponding to covert speech. Figure 44 shows the average power change during covert speech compared to silence for the different electrodes and frequency bands. We observed that the considered band powers are overall weakly modulated. The most modulated electrodes are primarily located in the frontal areas. In particular, we could distinguish negative modulations of the 0-1 Hz band in pars opercularis and the premotor cortex as well as positive modulations in the supramarginal gyrus. Positive power changes of the bands covering beta rhythm (15-20 and 20-30 Hz) were observed on sites spanning from pars opercularis to the primary somatosensory cortex. No clear power change was observed in the low and high gamma frequency bands.

7.3.2 Offline open-loop decoding

We attempted the decoding of covertly produced phonemes based on the neural activity of a participant implanted with a 72-channel ECoG grid. The dataset that was used contained 3 repetitions of 21 phonemes. In each trial, the participant imagined continuously (and not repetitively) pronouncing each phoneme for 5 s. For each trial, only the 3-s interval in the middle of the covert production period was considered. The objectives were 1) to determine if the neural features allowed to decode acoustic and articulatory characteristics of the the imagined phonemes and 2) to determine which neural features were most contributing to this decoding.

For each of the 71 selected channels of the ECoG grid, the variations of power in the following frequency bands were computed: 0-1, 1-5, 5-10, 10-15, 15-20, 20-30, 30-40,

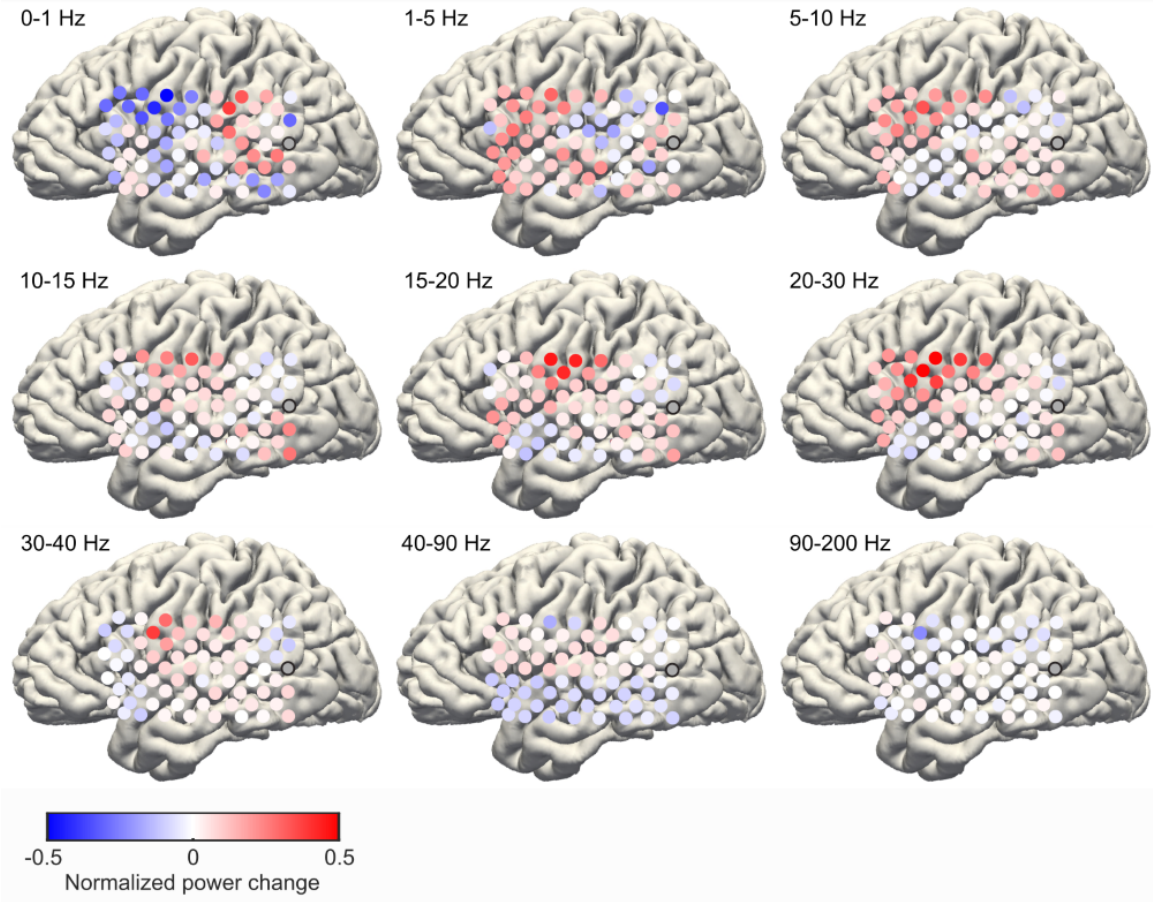


Figure 44. Band power modulation during covert speech. The normalized power change was computed by z-scoring the band powers using the mean and standard deviation computed from the silence-labeled intervals. For each electrode and frequency band, the obtained result was then averaged over all covert-labeled intervals. The extrema of the color scale were manually fixed to improve visibility, the values actually ranged from -0.7 to 0.7. The electrode represented by the grey dot circled in black was exclude from the analysis.

40-90, 90-200 Hz. These neural features, sampled at 100 Hz, were smoothed over 200-ms windows to improve their stability.

Typical acoustic and articulatory characteristics of the produced phonemes were associated to each sample of each trial, based on the data of another speaker.The acoustic representation consisted in 25 mel-cepstral filter coefficients (MELs), which represent the spectral envelope of the speech sounds. The articulatory representation consisted in 14 spatial coordinates, which describe the positions of 7 relevant points of the vocal tract.

Acoustic and articulatory parameters were decoded from the neural features using linear regression. Speech parameters were decoded for all trials in a LOO cross-validation. In each fold, the N_s best neural features were determined and a linear model was trained. The whole process was carried out for different values of N_s. Except in section 7.3.2.1 where the influence of N_s is studied, N_s was optimized for the results that are displayed and discussed below.

7.3.2.1 Influence of features selection

For each decoding, the number of features used in the regression model was limited to N_s. The decoding performance was evaluated for N_s taking values between 10 and 200. In each case, the N_s best features were determined as those which contributed the most to a linear model built using all features. As explained below, the best performance was obtained for the decoding of acoustic parameters using 71 neural features: the 0-1 Hz band powers for the 71 electrodes. For this decoding the optimal value of N_s was 40 (Figure 45a). To evaluate the influence of the feature selection, we compared the case of $N_s = 40$ to the case of $N_s = 70$. The case $N_s = 70$ was assimilated as an absence of feature selection. Selecting 40 features instead 70 increased the average correlation from 0.17 ($\pm$ 0.1) to 0.22 ($\pm$ 0.09). This increase was significantly superior to zero (one-tailed paired-sample Student's t-test, $p < 0.05$), which shows that the optimization of the number of selected features was beneficial to the decoding performance.

For the other decoding results exposed in the following sections, different values of N_s were also tested but only the best result was considered.

7.3.2.2 Decoding of acoustic and articulatory representations

We attempted to decode acoustic and articulatory representations of the imagined phonemes based on different subsets of the neural features. The decoding performance was evaluated by computing the correlations between the decoded and target speech representation (Figure 45b).

In the best case, the decoding of acoustic coefficients reached an average correlation of 0.22 ($\pm$ 0.09). The correlations obtained for the 25 MEL coefficients were all positive and ranged from 0.06 to 0.36 (Figure 45c). The correlations were statistically superior to chance level for 18 out of the 25 coefficients (permutation test, $p < 0.05$).

On the other hand, articulatory positions could only be decoded with average correlation coefficients inferior to 0.01. We concluded that the predicted articulatory positions were not correlated to the expected ones. In consequence, only the acoustic decoding results are detailed in the following sections.

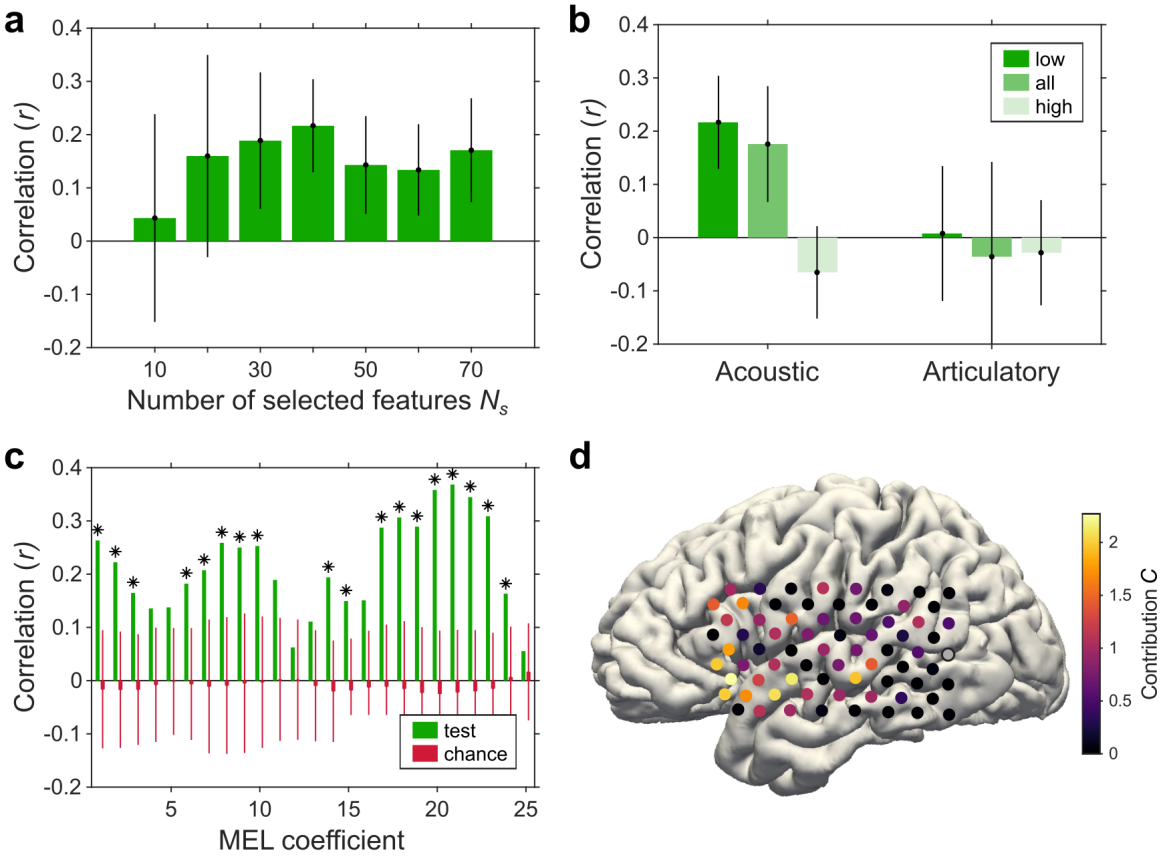


Figure 45. Results of the offline decoding of covert phonemes. (a) Average correlation of the MEL coefficients decoded using different numbers of features N_s, selected among the 0-1 Hz LFP band powers. The optimal value of N_s was 40. **(b)** Average correlations of decoded acoustic coefficients and articulatory positions using subsets of the neural features. The 3 subsets correspond to subsets of LFP frequency band powers: the 0-1 Hz band (low), the bands above 1 Hz (high) and the combination of the two (all). In each case, the number of selected features was optimized. **(c)** Correlations of the MEL coefficients decoding using the 0-1 Hz LFP band powers. The optimal number of features was used. The decoding process was applied 100 times with shuffled phonemes to estimate the chance levels. The stars indicate the correlations that were statistically superior to chance level (permutation test, $p < 0.05$). **(d)** Mapping on participant P5's brain of the contribution C of the electrodes to the decoding of the acoustic representation using 0-1 Hz band powers. The electrode represented by a circled grey dot represents the excluded electrode.

7.3.2.3 Contribution of the different frequency band powers of the LFP

The decoding of the acoustic representations was carried out using 3 different subsets of neural features. The first subset contained the frequency band powers of the 71 electrodes for the 0-1 Hz band. The second subset contained all the neural features. The third subset contained the frequency band powers for all the bands except the 0-1 Hz band. Separate decodings were computed using the 3 subsets. In each case, the number of selected features N_s was optimized.

The highest correlation was obtained for the first subset, using only the 0-1 Hz band powers ($r = 0.22 \pm 0.09$). The inclusion of the neural activity in the higher frequencies appeared to decrease the decoding performance ($r = 0.17 \pm 0.11$). Using only the frequency bands above 1 Hz, the obtained average correlation was slightly negative ($r = 0.07 \pm 0.09$). We conclude that the neural features contributing the most to the decoding are the 0-1 Hz band powers. The fact that the inclusion of the other frequency bands decreases the decoding performance could be explain by the fact that these additional features did not contribute to the decoding but increased the number of input features, which favors overfitting (see section 4.2).

7.3.2.4 Contribution of the electrodes

We computed the decoding contribution C of the neural features to the acoustic decoding based on 0-1 Hz band powers, using the same formula as for feature selection (see section 7.2.5.6). Since only one frequency band was used, the values of C represent the contribution of the different electrodes. When mapping the decoding contribution on the reconstruction of the participant's brain (Figure 45d), we observed that the most contributing electrodes were located on Broca's area and the superior temporal gyrus. More specifically, most relevant electrodes were clustered in an area comprising pars orbitalis, the ventral part of pars triangularis and the anterior part of the superior temporal gyrus. Interestingly, no clear modulation of the 0-1 Hz band power was observed for these electrodes when comparing covert speech and silence (Figure 44).

7.3.2.5 Decoding using low-passed LFP amplitude

We also attempted to decode the acoustic speech representation using the amplitude of the LFP low-passed below 1 Hz. The MEL coefficients were decoded with an average correlation of 0.04 ($\pm$ 0.06), much lower than when using the 0-1 Hz band powers.

7.3.2.6 Phoneme discrimination based on the decoded acoustic representations

Another way to evaluate the performance of the decoded acoustic representations is to quantify the ability to discriminate the different phonemes based on it. Figure 46 shows examples of target and decoded MELs resulting from the decoding based on the 0-1 Hz band powers. On these selected examples, it appears that the decoded MELs seemed close enough to the target to discriminate the different phonemes, at least on part of the trials. In order to quantitatively evaluate this aspect, we classified the decoded representations by matching each sample to the closest typical phoneme representation. To estimate the

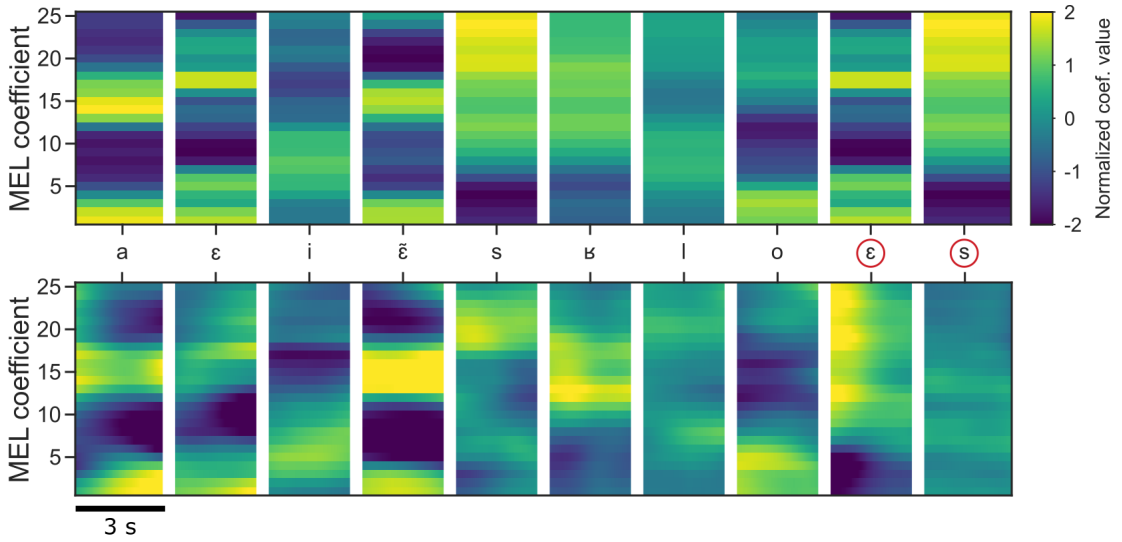


Figure 46. Examples of decoded acoustic coefficients. The upper panel shows the target values of the 25 MEL coefficients for 10 trials. For each trial, the target values were constant during the 3 seconds of covert speech production. The target values correspond to the typical acoustic representation of the requested phoneme, displayed between the two panels. The lower panel shows, for the 10 trials, the MELs resulting from the decoding based on the 0-1 Hz band powers. The first 8 trials were selected as examples of well decoded coefficients while the last 2 trials (circled in red) show examples of poor decoding.

chance level, the same process was repeated using the decoded representations obtained from the shuffled data. 9.5% of the decoded samples were correctly classified among the 21 phonemes. This result was significantly over the chance level of 4.6 $\pm$ 1.8% (permutation test, $p < 0.05$).

7.3.3 Online closed-loop decoding

In this experiment, the participant was asked to imagine pronouncing covert phonemes while her brain activity, recorded from a Utah array and an ECoG grid, was processed in real-time. A linear model, predicting articulatory positions from her brain activity, was continuously updated based on the data accumulating over the course of the session. An audio feedback was given to the subject during covert speech production. This feedback corresponded to the synthesis based on a weighted sum of the target and the decoded articulatory positions. At the beginning of each session, the synthesis was only based on the target positions and this assistance was gradually decreased.

The approach used in the online decoding, in particular the choice of decoding articulatory parameters, appears to be incoherent with offline decoding results. This is because both recordings were carried out during the same week. As a consequence, the offline decoding results were obtained after the closed-loop experiment.

7.3.3.1 Technical performances

The full processing chain of the closed-loop decoder was able to produce synthesized speech from the decoding of neural activity samples in less than 20 ms. This latency was satisfactory as it was shown that delays on audio feedback inferior or equal to 50 ms do not disrupt speech production (Stuart et al., 2002). The phoneme sounds synthesized using the typical articulatory positions (for example when using full assistance) could be well recognized.

7.3.3.2 Decoding performances

Two different decoding runs were performed during the online closed-loop decoding session. In the first run, only the vowels were considered and λ was fixed to 0.1. Three trials of each phoneme were carried out. In the second run, the vowels and the unvoiced consonants were considered and λ was optimized. Two repetitions of each phoneme were carried out. In both runs, during the last trials, the assistance parameter α was gradually decreased from 1 to 0 in 0.1 increments.

In both runs, intelligible synthesis was obtained while the assistance was turned on. Some of the synthesized sounds could be discriminated with α as low as 0.1. However, in both runs, as data was accumulating and assistance was decreasing, the synthesized sounds tended to converge towards an average undefined voice sound. Without any assistance, no intelligible sound could be produced.

7.4 Discussion

Using the ECoG recording of one participant, we showed that it was possible to continuously decode the acoustic representation of imagined phonemes with an average correlation of 0.22, very significantly above the chance level. Using the decoded representation, it was possible to correctly identify the imagined phoneme among 21 possibilities for 9.5% of the samples. On the other hand, the decoding of the articulatory representation did not yield significant results. This suggests that the acoustic representation of speech might be better suited than the articulatory one to the decoding of covert speech.

We characterized the contribution of the different neural features based on the value of the associated coefficients in the linear regression. Using this criterion to select features from the training set had a significant positive influence on the decoding performance. It appeared that using only the 0-1 Hz band power of 40 electrodes yielded the best decoding performance. The most contributing electrodes were located in Broca's area and in the superior temporal gyrus. Surprisingly, these features did not appear to be modulated when comparing covert speech and silence. There was no evidence that frequency powers above 1 Hz could be used to decode the acoustic parameters, while modulation of the beta band power was on average observed over the inferior frontal gyrus and the sensorimotor cortex during covert speech. These findings suggest that power changes and other indicators of speech-related modulation might not always be helpful to select the features that are the most relevant to decode speech parameters. We also showed that the decoding performance was not statistically significant when using the LFP low-passed below 1 Hz

instead of the 0-1 Hz band power. Altogether, these results were unexpected as the 0-1 Hz band power has to our knowledge never been shown to be relevant to speech decoding. We hypothesize that the reason why slow power variations were relevant in our study might reside in the nature of the task performed by the participant. Indeed, the participant continuously imagined the same phoneme for 5 seconds. In this context, the slow power of the LFP might be more relevant than the other neural features to decode speech features that are stable over the course of whole trials.

The participant also took part in a closed-loop decoding experiment. Our implementation gave very satisfactory results on the technical side, as we were able to extract neural features from the signals of a 72-electrode ECoG grid and a 96-electrode Utah array, perform linear regression and synthesize speech, with a total processing time inferior to 20 ms. However, we could not show that the articulatory features were correctly decoded as the synthesized sounds did not match the targets when no assistance was used.

Several factors, discovered in retrospective analyses, might explain this absence of clear result. According to shape of their waveforms, it is likely that the putative action potentials were in fact high-frequency noise. First, we could not confirm that action potentials have been recorded by the Utah array. Feeding the decoder with irrelevant features could have negatively affected the decoding performance. Secondly, one noisy ECoG channel containing high amplitude noise was included in the CAR, which might have degraded the overall quality of the signals. Thirdly, since the data used by the online decoder was similar to what was used offline (same participant, similar task and similar amount of training data), we suppose that using the 0-1 Hz band power, decoding acoustic instead of articulatory features and using feature selection could have yielded better results.

Finally, we retrospectively found that acoustic contamination caused by the audio feedback was observed in the ECoG recording above 100 Hz, as visible in Figure 14. The contamination could have positively affected the decoding performance. We therefore think that the presence of acoustic contamination should have been assessed during the experiment and that the contaminated frequencies should have been excluded from the neural features.

To conclude, we found that the 0-1 Hz band powers in Broca's area and the superior temporal gyrus could be used to decode sustained covert phoneme production. Modulations of this particular feature has not been reported. It is interesting to relate these low-frequency power variations to the modulations of the low-passed LFP observed in Broca's area in Chapter 6. These two measurements differ on several points (epicortical vs. intracortical, macroscopic vs. microscopic, power vs. amplitude), but could be related to a common phenomenon. Observing these slow modulations in Broca's area during other speech-related tasks and in different subjects is necessary to better understand it. It seems however unlikely that such intrinsically slow-varying features would be relevant to the decoding of more naturally paced speech, which implies that sustained covert speech might not be a suitable way to calibrate a closed-loop BCI intended for this purpose.

Conclusions and perspectives

The aim of the present thesis was to contribute to the development of speech BCIs by tackling some of the challenges faced in this research field.

First of all, we exposed the problem of acoustic contamination of electrophysiological signals, which we showed could bias some of the current and future speech decoding studies. Secondly, we hypothesized that decoding the neural activity elicited by attempted speech in people unable to speak might require different decoding strategies (different neural features, speech features, recording sites) than for overt speech. To get a better sense of the cortical activity in the absence of sensory feedbacks, we decided to study covert speech in healthy participants. To obtain new insights on the different brain dynamics during overt and covert speech, we analyzed the activity of a population of neurons in Broca's area during these two conditions. Finally, we attempted to decode covert speech features from electrocorticographic data using a calibration protocol that could theoretically be used with a person unable to speak.

In a first part, we highlighted the existence of the phenomenon of acoustic contamination of electrophysiological signals using human and animal recordings during speech production and/or sound perception experiments. We found that in some recordings, unexpectedly high correlations were observed between the spectrograms of the electrophysiological signals and the spectrogram of the simultaneously recorded audio signal. We confirmed that these correlations were not imputable to brain activity by reproducing the phenomenon in a minimal in vitro setup. Additional analyses revealed that the contamination was likely to be caused by a microphonic effect in the acquisition chain, due to the vibration of connectors. We found that acoustic contamination is a widespread phenomenon, as it affected recordings from 3 out 5 research teams who participated in our study. Acoustic contamination is particularly problematic for studies attempting to decode speech from neural activity during production or perception. Indeed, the high-gamma frequency band power was often used as a neural feature in recent decoding studies and can be contaminated by the participant's voice or the speech stimuli. Acoustic contamination could also potentially bias the detection of multi-unit activity as it affects the corresponding frequency range, but this remains to be precisely investigated.

Our results alert on the fact that acoustic contamination could bias the results of studies in which the electrophyiological signals are acquired in the presence of sound, in particular speech decoding studies. In order to avoid acoustic contamination in the future, it would be interesting to confirm that a microphonic effect occurs only at the connector levels and to solve this problem by designing sound-proof connection systems. In the meantime, we suggest that researchers in this field should systematically assess

the presence or absence of acoustic contamination in the neural data. To this end, we have proposed methods to visualize and statistically assess the presence of acoustic contamination. Adapting this methods to run real-time diagnoses could prove useful to detect potential contamination while conducting experiments. Indeed, in some cases it might be possible to suppress or reduce the contamination by identifying loose connectors or isolating the acquisition system from the sound source. Following our publication on this topic (Roussel et al., 2020), recent articles and pre-prints in the field of speech decoding have started to assess the potential presence of acoustic contamination in their data (Wilson et al., 2020; Angrick et al., 2020; Proix et al., 2021). In particular, Wilson et al. (2020) proposed to use linear regression reference (LRR; Young et al., 2018) as a post-processing method to mitigate acoustic contamination. They showed that LRR seemed to partially remove acoustic contamination and that doing so decreased the performance of the decoder. These results support the idea that acoustic contamination biases decoding performances.

In a second part, based on an intracortical micro-electrode recording, we were able, for the first time, to study the activity of a neuronal population of Broca's area during over and covert speech. Our analysis revealed that the firing rates of this population allowed to discriminate if the participant was overtly speaking, covertly speaking or not speaking. Using the the Victor and Purpura distance, we further evaluated the similarity between pairs of spike trains at the population level. We found that pairs of spike trains elicited by two overt repetitions of the same sentence tended to be more similar than pairs of spike trains elicited by the utterances of two different sentences. The similarity was high enough to classify 31 sentences based on spike trains over chance level. We also showed that the multi-unit spike train similarity was correlated to the phonological similarity of sentences. By contrast, we could not find such correlation during covert repetitions, which supports the idea that the neuronal dynamics are different compared to overt speech. Finally, we observed speech-related modulations of the low-frequency LFP (below 1 Hz). On part of the electrodes, positive or negative peaks were observed during overt speech. The polarization pattern was partially similar in the covert speech condition, but only visible on a few electrodes. For a majority of units, action potentials were preferentially emitted during the troughs of the low-passed LFP.

The results of our analyses on spiking data and low-frequency LFP show that modulations of neural activity are observed in Broca's area during speech production, which contrasts with ECoG studies that have shown that this region was mostly active before and not during speech (Flinker et al., 2015; Brumberg et al., 2016). Our observations also showed that the spiking and LFP activities in Broca's area are different in overt and covert speech. The analysis of the firing rates showed that this difference is not just simply a reduced spiking activity in covert speech. The fact that spike distances were correlated to the phonological content in overt but not in covert speech supports the idea that the spiking patterns are different. In consequence overt speech decoders based on the activity of Broca's area during overt speech might not transfer well to the decoding of covert speech. In the case of overt speech, the fact that no rate coding has been observed at the level of individual neurons and that the spike train similarity was correlated to the phonological content suggest that phonemes, or group of phonemes, could be encoded in spike timings in Broca's area. Additional neuronal recordings in this area during different

speech-related tasks would allow to explore further the nature of the encoding of the phonological content. In particular, given the hypothesized role of Broca's area in handling phonological representations, it would be interesting to analyze the spiking patterns elicited by isolated phonemes as well as simple combinations of these phonemes. Altogether, the present results do not show Broca's area as a particularly relevant recording region for intracortical speech BCIs. Indeed, if speech features are actually encoded by the spike timing of neurons with low firing rates, efficient decoding would likely require to record hundreds of individual units in a very stable way over months, which is currently a technically challenging task. New advances on neurotechnologies, such as the development of implants with thousands of micro-electrodes (Musk and Neuralink, 2019), might open new perspectives on such topics in the next years.

In a third part, we attempted to decode acoustic and articulatory features of imagined phonemes based on the electrophysiological recordings of one participant. The participant was asked to imagine pronouncing a phoneme during several seconds. Using this protocol of sustained production allowed us to circumvent the problem of incertitude on the covert speech timing. We estimated articulatory and acoustic representations of the imagined phonemes based on the data of another speaker. We then used the neural features to continuously decode these speech features over the course of each trial. Using the data of an open-loop recording, we could decode the articulatory representation with a statistically significant performance. This allowed to correctly identify the imagined phoneme for about 10% of the neural data samples, statistically above chance level. By contrast, decoding articulatory representation led to poor decoding performances. The best acoustic decoding was obtained using only the spectral power in the 0-1 Hz frequency band, with the most contributing electrodes being located in Broca's area and the superior temporal gyrus. The same participant also took part in a closed-loop paradigm during which a similar decoder was used and real-time audio feedback was provided to the user. In this experiment, which took place before obtaining the results exposed above, the articulatory positions were used as the target representation of the decoder. On a technical point of view, this first attempt was satisfactory as it was possible to process the raw signals, decode articulatory positions and synthesize sound in real-time with a very acceptable delay. On the neuroscientific aspect however, we could not obtain clear evidence that the articulatory positions were correctly decoded. We retrospectively identified several problems in the acquisition process that might have negatively affected the decoding.

The results of the offline decoding showed that the 0-1 Hz band powers in Broca's area and the superior temporal gyrus were the most informative features to decode sustained covert phoneme production. Modulations of these very-low-frequency components of the LFP had, to our knowledge, never been reported in relation to speech production. This result echoes to the slow variations of potential that we observed in Broca's area during speech in the intracortical recording of participant P3 – but it remains for now difficult to know if they are related. It is also unclear why the frequency bands that have traditionally been used for speech decoding did not appear to be relevant in our study. One of the hypothesis is that the neural activity elicited by the chosen task, i.e. sustained phoneme production, might not be representative of the activity underlying more naturally paced speech. If that is the case, then it implies that the features fed to the online decoder, should ideally also have been the 0-1 Hz band powers. However,

considering the potential use of low-frequency power variations in speech BCIs, it seems unlikely that such intrinsically slow-varying features would be relevant to the decoding of naturally paced speech. This implies that a different protocol, involving tasks that are closer to natural speech, would probably be better suited to the calibration of a BCI based on covert speech. A possible protocol, which also handles the problem of incertitude on covert speech timing, could be to ask the participant to imagine pronouncing sounds/words at a slow pace, in synchrony with a visual stimulus. This protocol could be adapted to a closed-loop calibration by using the same principle of assistance as in our first attempt. In that case the audio feedback could also be generated by a weighted sum of the target and the decoded speech features.

The results of this thesis confirm that difficult challenges remain to be faced to decode speech in real-time from the brain activity of people that are unable to speak due to paralysis. If the performance obtained on overt speech decoding keep improving, successfully decoding covert speech might be the next major step towards this goal. However, drawing a parallel with motor tasks invite us to consider that the activity underlying covert speech in healthy participants might not be completely identical to the one observed in paralyzed people attempting to speak. Indeed, it has been shown in fMRI imaging studies that attempted motor execution and motor imagery do not elicit exactly the same cortical activation patterns in paralyzed individuals (Hotz-Boendermaker et al., 2008; Szameitat et al., 2012). A study of multi-unit activity in the hand motor area also showed that stronger neuronal activation is observed during attempted motor execution (Rastogi et al., 2020). Even though covert speech in healthy participants is a useful proxy to develop speech decoders that could be used by paralyzed individuals, these results suggest that it may be limited. In consequence, studying the difference of cortical activation during speech imagery and attempt in paralyzed subjects could help to assess these limits and guide future choices of decoding strategy.

Annexes

Annex 1: International Phonetic Alphabet

All the phonetic transcriptions in this thesis use the International Phonetic Alphabet (IPA). The following page displays the IPA chart, available on the website of the International Phonetic Association[1] (under Creative Commons Attribution-Sharealike 3.0 Unported License, Copyright © 2020 International Phonetic Association).

[1] Available at http://www.internationalphoneticassociation.org/content/ipa-chart

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	’ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives

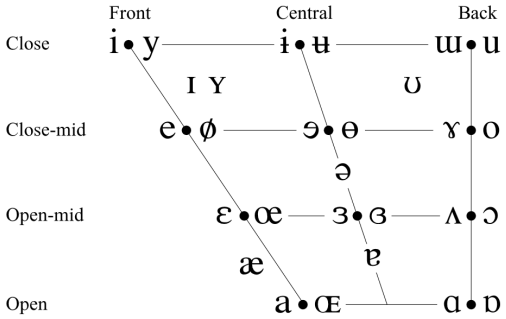
ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

t͡s k͡p

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ Primary stress ˌfoʊnəˈtɪʃən

ˌ Secondary stress

ː Long eː

ˑ Half-long eˑ

˘ Extra-short ĕ

| Minor (foot) group

‖ Major (intonation) group

. Syllable break ɹi.ækt

‿ Linking (absence of a break)

DIACRITICS

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	~	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝ (ɪ̝ = voiced alveolar fricative)			
◌̩	Syllabic	n̩	◌̞	Lowered	e̞ (β̞ = voiced bilabial approximant)			
◌̯	Non-syllabic	e̯	◌̙	Advanced Tongue Root	e̙			
◌˞	Rhoticity	ə˞ a˞	◌̘	Retracted Tongue Root	e̘			

Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

TONES AND WORD ACCENTS

LEVEL		CONTOUR	
e̋ or e᷅	Extra high	ě or e᷄	Rising
é	High	ê	Falling
ē	Mid	e᷌	High rising
è	Low	e᷈	Low rising
ȅ	Extra low	e᷉	Rising-falling
↓	Downstep	↗	Global rise
↑	Upstep	↘	Global fall

Bibliography

Abdoun, O. NeuroMap: a spline-based interactive open-source software for spatiotemporal mapping of 2D and 3D MEA data. *Frontiers in Neuroinformatics*, 4:119, jan 2011.

Ackermann, H. and Riecker, A. The contribution of the insula to motor aspects of speech production: A review and a hypothesis. *Brain and Language*, 89(2):320–328, may 2004.

Affolter, N., Egressy, B., Pascual, D., and Wattenhofer, R. Brain2Word: Decoding Brain Activity for Language Generation. *arXiv*, sep 2020.

Akbari, H., Khalighinejad, B., Herrero, J., Mehta, A., and Mesgarani, N. Towards reconstructing intelligible speech from the human auditory cortex. *Nature Scientific Reports*, page 350124, 2019.

Amunts, K., Lenzen, M., Friederici, A. D., Schleicher, A., Morosan, P., Palomero-Gallagher, N., and Zilles, K. Broca's region: Novel organizational principles and Multiple Receptor Mapping. *PLoS Biology*, 8(9):e1000489, sep 2010.

Angrick, M., Ottenhoff, M., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., Saal, J., Colon, A. J., Wagner, L., Krusienski, D. J., Kubben, P. L., Schultz, T., and Herff, C. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity, dec 2020. ISSN 26928205. URL `https://doi.org/10.1101/2020.12.11.421149`.

Anumanchipalli, G. K., Chartier, J., and Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

Aziz-Zadeh, L., Cattaneo, L., Rochat, M., and Rizzolatti, G. Covert speech arrest induced by rTMS over both motor and nonmotor left hemisphere frontal sites. *Journal of Cognitive Neuroscience*, 17(6):928–938, jun 2005.

Benabid, A. L., Costecalde, T., Eliseyev, A., Charvet, G., Verney, A., Karakas, S., Foerster, M., Lambert, A., Morinière, B., Abroug, N., Schaeffer, M. C., Moly, A., Sauter-Starace, F., Ratel, D., Moro, C., Torres-Martinez, N., Langar, L., Oddoux, M., Polosan, M., Pezzani, S., Auboiroux, V., Aksenova, T., Mestais, C., and Chabardes, S. An exoskeleton controlled by an epidural wireless brain–machine interface in a tetraplegic patient: a proof-of-concept demonstration. *The Lancet Neurology*, 18(12):1112–1122, dec 2019.

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A., and Ramsey, N. F. Neural tuning to low-level features of speech throughout the perisylvian cortex. *Journal of Neuroscience*, 37(33):7906–7920, 2017.

Bocquelet, F. *Toward a brain-computer interface for speech restoration*. PhD thesis, Université Grenoble Alpes, apr 2017.

Bocquelet, F., Hueber, T., Girin, L., Chabardès, S., and Yvert, B. Key considerations in designing a speech brain-computer interface. *Journal of Physiology Paris*, 110(4): 392–401, 2016a.

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., and Yvert, B. Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces. *PLoS Computational Biology*, 12(11), 2016b.

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., and Yvert, B. Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces. *PLoS Computational Biology*, 12(11):1–28, 2016c.

Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, 2013.

Bouchard, K. E., Conant, D. F., Anumanchipalli, G. K., Dichter, B., Chaisanguanthum, K. S., Johnson, K., and Chang, E. F. High-resolution, non-invasive imaging of upper vocal tract articulators compatible with human brain recordings. *PLoS ONE*, 11(3): 1–30, 2016.

Brandman, D., Hosman, T., Saab, J., ..., Harrison, M., Simeral, J., and Hochberg, L. R. Rapid calibration of an intracortical brain computer interface for people with tetraplegia. *Journal of neural engineering*, (In press)(15), 2017.

Brumberg, J. S., Wright, E. J., Andreasen, D. S., Guenther, F. H., and Kennedy, P. R. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. *Frontiers in Neuroscience*, 5(MAY):1–12, may 2011.

Brumberg, J. S., Krusienski, D. J., Chakrabarti, S., Gunduz, A., Brunner, P., Ritaccio, A. L., and Schalk, G. Spatio-Temporal Progression of Cortical Activity Related to Continuous Overt and Covert Speech Production in a Reading Task. *PLOS ONE*, 11 (11):e0166872, nov 2016.

Carey, D., Krishnan, S., Callaghan, M. F., Sereno, M. I., and Dick, F. Functional and Quantitative MRI Mapping of Somatomotor Representations of Human Supralaryngeal Vocal Tract. *Cerebral cortex (New York, N.Y. : 1991)*, 27(1):265–278, jan 2017.

Cerf, M., Thiruvengadam, N., Mormann, F., Kraskov, A., Quiroga, R. Q., Koch, C., and Fried, I. On-line, voluntary control of human temporal lobe neurons. *Nature*, 467 (7319):1104–1108, oct 2010.

Chapin, J. K., Moxon, K. A., Markowitz, R. S., and Nicolelis, M. A. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience*, 2(7):664–670, 1999.

Chartier, J., Anumanchipalli, G. K., Johnson, K., and Chang, E. F. Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex. *Neuron*, 98(5): 1042–1054.e4, 2018.

Cheung, C., Hamiton, L. S., Johnson, K., and Chang, E. F. The auditory representation of speech sounds in human motor cortex. *eLife*, 5(MARCH2016):1–19, 2016.

Chi, T., Ru, P., and Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

Chicco, D. Ten quick tips for machine learning in computational biology, dec 2017. ISSN 17560381. URL `https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3`.

Chicco, D. and Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1): 6, jan 2020.

Christie, B. P., Tat, D. M., Irwin, Z. T., Gilja, V., Nuyujukian, P., Foster, J. D., Ryu, S. I., Shenoy, K. V., Thompson, D. E., and Chestek, C. A. Comparison of spike sorting and thresholding of voltage waveforms for intracortical brain-machine interface performance. *Journal of Neural Engineering*, 12(1), feb 2015.

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E. C., Weber, D. J., McMorland, A. J., Velliste, M., Boninger, M. L., and Schwartz, A. B. High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet*, 381 (9866):557–564, 2013.

Conant, D. F., Bouchard, K. E., Leonard, M. K., and Chang, E. F. Human sensorimotor cortex control of directly measured vocal tract movements during vowel production. *Journal of Neuroscience*, 38(12):2955–2966, 2018.

Corballis, M. C. Left Brain, Right Brain: Facts and Fantasies. *PLoS Biology*, 12(1), jan 2014.

Crone, N. E., Hao, L., Hart, J., Boatman, D., Lesser, R. P., Irizarry, R., and Gordon, B. Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology*, 57(11):2045–2053, dec 2001.

Damerau, F. J. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, mar 1964.

Dronkers, N. F. A new brain region for coordinating speech articulation. *Nature*, 384 (6605):159–161, 1996.

Elbert, T., Rockstroh, B., Lutzenberger, W., and Birbaumer, N. Biofeedback of slow cortical potentials. I. *Electroencephalography and Clinical Neurophysiology*, 48(3):293–301, mar 1980.

Eliseyev, A. and Aksenova, T. Penalized Multi-Way Partial Least Squares for Smooth Trajectory Decoding from Electrocorticographic (ECoG) Recording. *PLOS ONE*, 11 (5):e0154878, may 2016.

Engwall, O. Modeling of the vocal tract in three dimensions. In *EUROSPEECH'99*, pages 113–116, Budapest, Hungary, 1999.

Fairbanks, G. Systematic Research In Experimental Phonetics:* 1. A Theory Of The Speech Mechanism As A Servosystem. *The Journal of speech and hearing disorders*, 19(2):133–139, 1954.

Farouk, M. Measuring Sentences Similarity: A Survey. *arXiv*, 12:974–6846, oct 2019.

Farwell, L. A. and Donchin, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, dec 1988.

Fernyhough, C. Alien voices and inner dialogue: Towards a developmental account of auditory verbal hallucinations. *New Ideas in Psychology*, 22(1):49–68, apr 2004.

Fetz, E. E. Operant conditioning of cortical unit activity. *Science*, 163(3870):955–958, feb 1969.

Fetz, E. E. and Baker, M. A. Operantly conditioned patterns on precentral unit activity and correlated responses in adjacent cells and contralateral muscles. *Journal of neurophysiology*, 36(2):179–204, mar 1973.

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., and Crone, N. E. Redefining the role of broca's area in speech. *Proceedings of the National Academy of Sciences of the United States of America*, 112(9): 2871–2875, mar 2015.

Flint, R. D., Wright, Z. A., Scheid, M. R., and Slutzky, M. W. Long term, stable brain machine interface performance using local field potentials and multiunit spikes. *Journal of Neural Engineering*, 10(5):056005, oct 2013.

Foerster, O. and Penfield, W. The structural basis of traumatic epilepsy and results of radical operation. *Brain*, 53(2):99–119, jul 1930.

Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D. B., and Rorden, C. Revealing the dual streams of speech processing. *Proceedings of the National Academy of Sciences of the United States of America*, 113(52):15108–15113, 2016.

Gehrig, J., Michalareas, G., Forster, M. T., Lei, J., Hok, P., Laufs, H., Senft, C., Seifert, V., Schoffelen, J. M., Hanslmayr, S., and Kell, C. A. Low-Frequency Oscillations Code Speech during Verbal Working Memory. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 39(33):6498–6512, aug 2019.

Gershman, S. and Tenenbaum, J. Phrase similarity in humans and machines. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Austin, TX, 2015.

Geva, S., Jones, P. S., Crinion, J. T., Price, C. J., Baron, J.-C., and Warburton, E. A. The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. *Brain*, 134(10):3071–3082, oct 2011.

Gilja, V., Nuyujukian, P., Chestek, C. A., Cunningham, J. P., Yu, B. M., Fan, J. M., Churchland, M. M., Kaufman, M. T., Kao, J. C., Ryu, S. I., and Shenoy, K. V. A high-performance neural prosthesis enabled by control algorithm design. *Nature Neuroscience*, 15(12):1752–1757, dec 2012.

Goucha, T. and Friederici, A. D. The language skeleton after dissecting meaning: A functional segregation within Broca's Area. *NeuroImage*, 114:294–302, jul 2015.

Grabski, K., Lamalle, L., Vilain, C., Schwartz, J. L., Vallée, N., Tropres, I., Baciu, M., Le Bas, J. F., and Sato, M. Functional MRI assessment of orofacial articulators: Neural correlates of lip, jaw, larynx, and tongue movements. *Human Brain Mapping*, 33(10): 2306–2321, oct 2012.

Grewe, B. F., Langer, D., Kasper, H., Kampa, B. M., and Helmchen, F. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nature Methods*, 7(5):399–405, may 2010.

Guenther, F. H. and Vladusich, T. A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25(5):408–422, sep 2012.

Guenther, F. H., Hampson, M., and Johnson, D. A Theoretical Investigation of Reference Frames for the Planning of Speech Movements. *Psychological Review*, 105(4):611–633, 1998.

Guenther, F. H., Ghosh, S. S., and Tourville, J. a. Neural Modeling and Imaging of the Cortical Interactions Underlying Syllable Production. *Brain And Language*, 96:280–301, 2006.

Guenther, F. H., Brumberg, J. S., Joseph Wright, E., Nieto-Castanon, A., Tourville, J. A., Panko, M., Law, R., Siebert, S. A., Bartels, J. L., Andreasen, D. S., Ehirim, P., Mao, H., and Kennedy, P. R. A wireless brain-machine interface for real-time speech synthesis. *PLoS ONE*, 4(12), 2009.

Gunduz, A., Brunner, P., Sharma, M., Leuthardt, E. C., Ritaccio, A. L., Pesaran, B., and Schalk, G. Differential roles of high gamma and local motor potentials for movement preparation and execution. *Brain-Computer Interfaces*, 3(2):88–102, apr 2016.

Hagoort, P. and Indefrey, P. The Neurobiology of Language Beyond Single Words. *Annual Review of Neuroscience*, 37(1):347–362, jul 2014.

Hampel, F. R. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383, jun 1974.

Hansen, P. C. The L-Curve and its Use in the Numerical Treatment of Inverse Problems. In Johnston, P., editor, *Computational Inverse Problems in Electrocardiology*, volume 4, chapter The L-Curv, pages 119–142. WIT Press, 2000.

Harris, Z. S. Distributional Structure. *WORD*, 10(2-3):146–162, 1954.

Hecht, M., Hillemacher, T., Gräsel, E., Tigges, S., Winterholler, M., Heuss, D., Hilz, M. J., and Neundörfer, B. Subjective experience and coping in ALS. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, 3(4):225–231, 2002.

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., and Schultz, T. Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9(May):1–11, 2015.

Herff, C., Diener, L., Angrick, M., Mugler, E., Tate, M. C., Goldrick, M. A., Krusienski, D. J., Slutzky, M. W., and Schultz, T. Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices. *Frontiers in Neuroscience*, 13:1267, nov 2019.

Herff, C., Krusienski, D. J., and Kubben, P. The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions, feb 2020. ISSN 1662453X.

Hickok, G. Computational neuroanatomy of speech production, feb 2012. ISSN 1471003X.

Hickok, G. and Poeppel, D. Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, 4(4):131–138, apr 2000.

Hickok, G. and Poeppel, D. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99, may 2004.

Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, may 2007.

Hickok, G., Houde, J., and Rong, F. Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, 69(3):407–422, feb 2011.

Hinterberger, T., Neumann, N., Pham, M., Kübler, A., Grether, A., Hofmayer, N., Wilhelm, B., Flor, H., and Birbaumer, N. A multimodal brain-based feedback and communication system. *Experimental Brain Research*, 154(4):521–526, 2004.

Hochberg, L. R., Serruya, M. D., Friehs, G. M., Mukand, J. A., Saleh, M., Caplan, A. H., Branner, A., Chen, D., Penn, R. D., and Donoghue, J. P. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, jul 2006.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., Van Der Smagt, P., and Donoghue, J. P. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485 (7398):372–375, may 2012.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9 (8):1735–1780, nov 1997.

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., and Rybchenko, S. I. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication*, 55(1):22–32, jan 2013.

Hotz-Boendermaker, S., Funk, M., Summers, P., Brugger, P., Hepp-Reymond, M. C., Curt, A., and Kollias, S. S. Preservation of motor programs in paraplegics as demonstrated by attempted and imagined foot movements. *NeuroImage*, 39(1):383–394, jan 2008.

Houde, J. F. and Chang, E. F. The cortical computations underlying feedback control in vocal production. *Current Opinion in Neurobiology*, 33:174–181, aug 2015.

House, A. S. On Vowel Duration in English. *Journal of the Acoustical Society of America*, 33(9):1174–1178, 1961.

Huang, J., Carr, T. H., and Cao, Y. Comparing cortical activations for silent and overt speech using event-related fMRI. *Human Brain Mapping*, 15(1):39–53, 2002.

Hueber, T., Benaroya, E. L., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, 52(4):288–300, apr 2010.

Hueber, T., Bailly, G., and Denby, B. Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface. In *Annual Conference of the Inter-national Speech Communication Association (InterSpeech 2012)*, page Tue.P3c.01, Portland, United States, 2012.

Hughes, G. F. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, 36(6):2014–2026, feb 2016.

Ikeda, S., Shibata, T., Nakano, N., Okada, R., Tsuyuguchi, N., Ikeda, K., and Kato, A. Neural decoding of single vowels during covert articulation using electrocorticography. *Frontiers in Human Neuroscience*, 8(MAR):125, mar 2014.

Illa, A., Meenakshi G, N., and Ghosh, P. K. A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5075–5079. IEEE, mar 2017.

Imai, S. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 1, pages 93–96. IEEE, 1983.

Indefrey, P. and Levelt, W. J. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144, may 2004.

Jarosiewicz, B., Masse, N. Y., Bacher, D., Cash, S. S., Eskandar, E., Friehs, G., Donoghue, J. P., and Hochberg, L. R. Advantages of closed-loop calibration in intracortical brain-computer interfaces for people with tetraplegia. *Journal of Neural Engineering*, 10(4): 046012, 2013.

Jarosiewicz, B., Sarma, A. A., Bacher, D., Masse, N. Y., Simeral, J. D., Sorice, B., Oakley, E. M., Blabe, C., Pandarinath, C., Gilja, V., Cash, S. S., Eskandar, E. N., Friehs, G., Henderson, J. M., Shenoy, K. V., Donoghue, J. P., and Hochberg, L. R. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science Translational Medicine*, 7(313):313ra179–313ra179, 2015.

Jorfi, M., Skousen, J. L., Weder, C., and Capadona, J. R. Progress towards biocompatible intracortical microelectrodes for neural interfacing applications. *Journal of Neural Engineering*, 12(1):011001, feb 2015.

Jou, S. C., Schultz, T., Walliczek, M., Kraft, F., and Waibel, A. Towards continuous speech recognition using surface electromyography. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2, pages 573–576, 2006.

Käthner, I., Wriessnegger, S. C., Müller-Putz, G. R., Kübler, A., and Halder, S. Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain-computer interface. *Biological Psychology*, 102(1):118–129, oct 2014.

Kawahara, H. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2:1303–1306, 1997.

Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., and Greger, B. Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of Neural Engineering*, 7(5):056007, oct 2010.

Kello, C. T. and Plaut, D. C. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, 116(4):2354–2364, oct 2004.

Kenter, T. and De Rijke, M. Short text similarity with word embeddings. In *International Conference on Information and Knowledge Management, Proceedings*, volume 19-23-Oct-, pages 1411–1420, New York, NY, USA, oct 2015. Association for Computing Machinery.

Kenter, T., Borisov, A., and de Rijke, M. Siamese CBOW: Optimizing Word Embeddings for Sentence Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 2, pages 941–951, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics.

Kim, S. P., Simeral, J. D., Hochberg, L. R., Donoghue, J. P., Friehs, G. M., and Black, M. J. Multi-state decoding of point-and-click control signals from motor cortical activity in a human with tetraplegia. In *Proceedings of the 3rd International IEEE EMBS Conference on Neural Engineering*, pages 486–489, 2007.

Klatt, D. H. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3):971–995, mar 1980.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C. K. Demixed principal component analysis of neural population data. *eLife*, 5:e10989, apr 2016.

Leuthardt, E. C., Pei, X. M., Breshears, J., Gaona, C., Sharma, M., Freudenberg, Z., Barbour, D., and Schalk, G. Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task. *Frontiers in Human Neuroscience*, 6 (MAY 2012):99, 2012.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, jul 2013.

Liu, Y., Coon, W. G., De Pesters, A., Brunner, P., and Schalk, G. The effects of spatial filtering and artifacts on electrocorticographic signals. *Journal of Neural Engineering*, 12(5):056008, aug 2015.

Loew, L. M., Cohen, L. B., Salzberg, B. M., Obaid, A. L., and Bezanilla, F. Charge-shift probes of membrane potential. Characterization of aminostyrylpyridinium dyes on the squid giant axon. *Biophysical Journal*, 47(1):71–77, jan 1985.

Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., and Schalk, G. Electrocorticographic representations of segmental features in continuous speech. *Frontiers in Human Neuroscience*, 9(FEB):97, feb 2015.

Luna-Lozano, P. S. and Pallas-Areny, R. Microphonics in biopotential measurements with capacitive electrodes. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, 2010.

Maeda, S. Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model. In *Speech Production and Speech Modelling*, pages 131–149. Springer Netherlands, 1990.

Makin, J. G., Moses, D. A., and Chang, E. F. Machine translation of cortical activity to text with an encoder-decoder framework. *Nature Neuroscience*, 23(4):575–582, mar 2020.

Makkonen, T., Ruottinen, H., Puhto, R., Helminen, M., and Palmio, J. Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms. *International Journal of Language and Communication Disorders*, 53(2):385–392, 2018.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H. J., Crone, N. E., Rieger, J., Schalk, G., Knight, R. T., and Pasley, B. N. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, 7(MAY):1–15, 2014.

Martin, S., Brunner, P., Iturrate, I., Del Millán, J. R., Schalk, G., Knight, R. T., and Pasley, B. N. Word pair classification during imagined speech using direct brain recordings. *Scientific Reports*, 6(1):25803, sep 2016.

Martínez, J. and Quiroga, R. Q. Spike Sorting. In Quiroga, R. Q. and Panzeri, S., editors, *Principles of neural coding*, chapter Spike Sort, pages 61–74. CRC Press, Boca Raton, 2013.

Matchin, W. G. A neuronal retuning hypothesis of sentence-specificity in Broca's area. *Psychonomic Bulletin and Review*, 25(5):1682–1694, oct 2018.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, feb 2014.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv*, jan 2013.

Miller, K. J. A library of human electrocorticographic data and analyses. *Nature Human Behaviour*, 3(11):1225–1235, nov 2019.

Miller, K. J., Abel, T. J., Hebb, A. O., and Ojemann, J. G. Rapid online language mapping with electrocorticography: Clinical article. *Journal of Neurosurgery: Pediatrics*, 7(5): 482–490, may 2011.

Morise, M., Yokomori, F., and Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99D(7):1877–1884, jul 2016.

Moritz, C. T., Perlmutter, S. I., and Fetz, E. E. Direct control of paralysed muscles by cortical neurons. *Nature*, 456(7222):639–642, dec 2008.

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., Shih, J. J., Krusienski, D. J., and Slutzky, M. W. Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, 11(3):035015, jun 2014.

Mugler, E. M., Tate, M. C., Livescu, K., Templer, J. W., Goldrick, M. A., and Slutzky, M. W. Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri. *The Journal of Neuroscience*, 4653:1206–18, 2018.

Musk, E. and Neuralink. An integrated brain-machine interface platform with thousands of channels, jul 2019. ISSN 1438-8871. URL `https://doi.org/10.1101/703801`.

Nawrot, M., Aertsen, A., and Rotter, S. Single-trial estimation of neuronal firing rates: From single-neuron spike trains to population activity. *Journal of Neuroscience Methods*, 94(1):81–92, dec 1999.

Nicolai, E. N., Michelson, N. J., Settell, M. L., Hara, S. A., Trevathan, J. K., Asp, A. J., Stocking, K. C., Lujan, J. L., Kozai, T. D., and Ludwig, K. A. Design choices for next-generation neurotechnology can impact motion artifact in electrophysiological and fast-scan cyclic voltammetry measurements. *Micromachines*, 9(10), 2018.

Nijboer, F., Sellers, E. W., Mellinger, J., Jordan, M. A., Matuz, T., Furdea, A., Halder, S., Mochty, U., Krusienski, D. J., Vaughan, T. M., Wolpaw, J. R., Birbaumer, N., and Kübler, A. A P300-based brain-computer interface for people with amyotrophic lateral sclerosis. *Clinical Neurophysiology*, 119(8):1909–1916, aug 2008.

Nourski, K. V., Steinschneider, M., Rhone, A. E., Oya, H., Kawasaki, H., Howard, M. A., and McMurray, B. Sound identification in human auditory cortex: Differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. *Brain and Language*, 148:37–50, sep 2015.

O'Keefe, J. and Recce, M. L. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*, 3(3):317–330, 1993.

O'Shaughnessy, D. A study of French vowel and consonant durations. *Journal of Phonetics*, 9(4):385–406, 1981.

Palmer, E. D., Rosen, H. J., Ojemann, J. G., Buckner, R. L., Kelley, W. M., and Petersen, S. E. An event-related fMRI study of overt and covert word stem completion. *NeuroImage*, 14(1 I):182–193, jul 2001.

Pandarinath, C., Nuyujukian, P., Blabe, C. H., Sorice, B. L., Saab, J., Willett, F. R., Hochberg, L. R., Shenoy, K. V., and Henderson, J. M. High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife*, 6, feb 2017.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. a., Crone, N. E., Knight, R. T., and Chang, E. F. Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251, jan 2012.

Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of Neural Engineering*, 8(4):046028, 2011a.

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., and Schalk, G. Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *NeuroImage*, 54(4):2960–2972, feb 2011b.

Penfield, W. and Boldrey, E. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60(4):389–443, dec 1937.

Peng, H., Long, F., and Ding, C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, aug 2005.

Perkell, J. S. *A physiologically-oriented model of tongue activity in speech production.* PhD thesis, Massachusetts Institute of Technology, 1974.

Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciu, M., and Lœvenbruck, H. What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural Brain Research*, 261: 220–239, mar 2014.

Peterson, G. E. and Barney, H. L. Control Methods Used in a Study of the Vowels. *Journal of the Acoustical Society of America*, 24(2):175–184, mar 1952.

Pineda, J. A., Silverman, D. S., Vankov, A., and Hestenes, J. Learning to control brain rhythms: Making a brain-computer interface possible. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):181–184, 2003.

Price, C. J. The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, 1191(1):62–88, mar 2010.

Proix, T., Delgado Saa, J., Christen, A., Martin, S., Knight, R. T., Tian, X., Poeppel, D., Doyle, W. K., Arnal, L. H., Mégevand, P., and Giraud, A.-L. Imagined speech can be decoded from low-and cross-frequency 1 features in perceptual space. *bioRxiv*, page 2021.01.26.428315, jan 2021.

Ramsey, N. F., Salari, E., Aarnoutse, E. J., Vansteensel, M. J., Bleichner, M. G., and Freudenburg, Z. V. Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. *NeuroImage*, 180:301–311, oct 2018.

Rastogi, A., Vargas-Irwin, C. E., Willett, F. R., Abreu, J., Crowder, D. C., Murphy, B. A., Memberg, W. D., Miller, J. P., Sweet, J. A., Walter, B. L., Cash, S. S., Rezaii, P. G., Franco, B., Saab, J., Stavisky, S. D., Shenoy, K. V., Henderson, J. M., Hochberg, L. R., Kirsch, R. F., and Ajiboye, A. B. Neural Representation of Observed, Imagined, and Attempted Grasping Force in Motor Cortex of Individuals with Chronic Tetraplegia. *Scientific Reports*, 10(1):12, dec 2020.

Ray, S. and Maunsell, J. H. R. Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biology*, 9(4):e1000610, apr 2011.

Rogalsky, C. and Hickok, G. The role of Broca's area in sentence comprehension. *Journal of Cognitive Neuroscience*, 23(7):1664–1680, jul 2011.

Rolston, J. D., Gross, R. E., and Potter, S. M. Common median referencing for improved action potential detection with multielectrode arrays. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, pages 1604–1607. IEEE Computer Society, 2009.

Roussel, P., Le Godais, G., Bocquelet, F., Palma, M., Hongjie, J., Zhang, S., Giraud, A. L., Mégevand, P., Miller, K., Gehrig, J., Kell, C., Kahane, P., Chabardès, S., and Yvert, B. Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception. *Journal of Neural Engineering*, aug 2020.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., and Halgren, E. Sequential Processing of Lexical, Grammatical, and Phonological Information within Broca's Area. *Science*, 326(5951):445–449, oct 2009.

Salatino, J. W., Williams, J. J., Vazquez, A. L., Purcell, E. K., Eles, J. R., Michelson, N. J., Kozai, T. D. Y., and Cui, X. T. Multi-scale, multi-modal analysis uncovers complex relationship at the brain tissue-implant neural interface: new emphasis on the biological interface. *Journal of Neural Engineering*, 15(3):033001, 2017.

Saura, D., Kreher, B. W., Schnell, S., Kümmerera, D., Kellmeyera, P., Vrya, M. S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., and Weiller, C. Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(46):18035–18040, nov 2008.

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, jun 2004.

Sellers, E. W., Krusienski, D. J., McFarland, D. J., and Wolpaw, J. R. Noninvasive Brain-Computer Interface Research at the Wadsworth Center. In Dornhege, G., Millan, J., Hinterberger, T., McFarland, D., and Müller, K., editors, *Toward Brain-Computer Interfacing*, chapter 2, pages 31–42. The MIT Press, Cambridge, MA, 2007.

Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R., and Donoghue, J. P. Instant neural control of a movement signal. *Nature*, 416(6877):141–142, mar 2002.

Small, L. H. *Fundamentals of phonetics: a practical guide for students*. Pearson, 3rd edition, 2012.

Song, C., Xu, R., and Hong, B. Decoding of Chinese phoneme clusters using ECoG. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, pages 1278–1281. IEEE, aug 2014.

SPTK Working Group. Speech Signal Processing Toolkit (SPTK), 2017. URL `http://sp-tk.sourceforge.net/`.

Stavisky, S. D., Kao, J. C., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. A high performing brain-machine interface driven by low-frequency local field potentials alone and together with spikes. *Journal of Neural Engineering*, 12(3):036009, jun 2015.

Stuart, A., Kalinowski, J., Rastatter, M. P., and Lynch, K. Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America*, 111(5):2237, 2002.

Szameitat, A. J., Shen, S., Conforto, A., and Sterr, A. Cortical activation during executed, imagined, observed, and passive wrist movements in healthy volunteers and stroke patients. *NeuroImage*, 62(1):266–280, aug 2012.

Tankus, A., Fried, I., and Shoham, S. Structured neuronal encoding and decoding of human speech features. *Nature Communications*, 3(1):1015, jan 2012.

Tate, M. C., Herbet, G., Moritz-Gasser, S., Tate, J. E., and Duffau, H. Probabilistic map of critical functional regions of the human cerebral cortex: Broca's area revisited. *Brain*, 137(10):2773–2782, 2014.

Toda, T., Black, A. W., and Tokuda, K. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3):215–227, mar 2008.

Torre, E., Quaglio, P., Denker, M., Brochier, T., Riehle, A., and Grün, S. Synchronous spike patterns in macaque motor cortex during an instructed-delay reach-to-grasp task. *Journal of Neuroscience*, 36(32):8329–8340, 2016.

Tourville, J. A., Nieto-Castañón, A., Heyne, M., and Guenther, F. H. Functional Parcellation of the Speech Production Cortex. *Journal of Speech, Language, and Hearing Research*, 62(8S):3055–3070, aug 2019.

Townsend, G. and Platsko, V. Pushing the P300-based brain-computer interface beyond 100 bpm: Extending performance guided constraints into the temporal domain. *Journal of Neural Engineering*, 13(2):026024, apr 2016.

Toyoda, G., Brown, E. C., Matsuzaki, N., Kojima, K., Nishida, M., and Asano, E. Electrocorticographic correlates of overt articulation of 44 English phonemes: Intracranial recording in children with focal epilepsy. *Clinical Neurophysiology*, 125(6):1129–1137, jun 2014.

Tremblay, P. and Dick, A. S. Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language*, 162:60–71, nov 2016.

Trupe, L. A., Varma, D. D., Gomez, Y., Race, D., Leigh, R., Hillis, A. E., and Gottesman, R. F. Chronic apraxia of speech and Broca's area. *Stroke*, 44(3):740–744, mar 2013.

Umeda, N. Consonant duration in American English. *Journal of the Acoustical Society of America*, 61(3):846–858, 1977.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv*, sep 2016.

Van Der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Ventura, M. I., Nagarajan, S. S., and Houde, J. F. Speech target modulates speaking induced suppression in auditory cortex. *BMC Neuroscience*, 10(1):58, jun 2009.

Victor, J. D. and Purpura, K. P. Nature and precision of temporal coding in visual cortex: A metric- space analysis. *Journal of Neurophysiology*, 76(2):1310–1326, 1996.

Victor, J. D. and Purpura, K. P. Metric-space analysis of spike trains: theory, algorithms and application. *Network: Computation in Neural Systems*, 8(2):127–164, 1997.

Vigneau, M., Beaucousin, V., Hervé, P. Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., and Tzourio-Mazoyer, N. Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage*, 30(4):1414–1432, may 2006.

Wessberg, J., Stambaugh, C. R., Kralik, J. D., Beck, P. D., Laubach, M., Chapin, J. K., Kim, J., Biggs, S. J., Srinivasan, M. A., and Nicolelis, M. A. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408(6810): 361–365, nov 2000.

Wilson, G. H., Stavisky, S. D., Willett, F. R., Avansino, D. T., Kelemen, J. N., Hochberg, L. R., Henderson, J. M., Druckmann, S., and Shenoy, K. V. Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. *Journal of neural engineering*, 17(6):066007, 2020.

Wodlinger, B., Downey, J. E., Tyler-Kabara, E. C., Schwartz, A. B., Boninger, M. L., and Collinger, J. L. Ten-dimensional anthropomorphic arm control in a human brain-machine interface: Difficulties, solutions, and limitations. *Journal of Neural Engineering*, 12(1): 16011, feb 2015.

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., and Vaughan, T. M. Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*, 8(2):164–173, jun 2000.

Wu, W., Gao, Y., Bienenstock, E., Donoghue, J. P., and Black, M. J. Bayesian population decoding of motor cortical activity using a Kalman filter. *Neural Computation*, 18(1): 80–118, jan 2006.

Yaksi, E. and Friedrich, R. W. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. *Nature Methods*, 3(5):377–383, may 2006.

Yang, X., Wang, K., and Shamma, S. A. Auditory Representations of Acoustic Signals. *IEEE Transactions on Information Theory*, 38(2):824–839, 1992.

Young, D., Willett, F., Memberg, W. D., Murphy, B., Walter, B., Sweet, J., Miller, J., Hochberg, L. R., Kirsch, R. F., and Ajiboye, A. B. Signal processing methods for reducing artifacts in microelectrode brain recordings caused by functional electrical stimulation. *Journal of Neural Engineering*, 15(2), jan 2018.

Ze, H., Senior, A., and Schuster, M. Statistical parametric speech synthesis using deep neural networks. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 7962–7966, oct 2013.

Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., and Cichocki, A. Higher order partial least squares (HOPLS): A generalized multilinear regression method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1660–1673, 2013.