Decoding speech from brain activity using linear methods

Gaël Le Godais

Thèse pour obtenir le grade de

Docteur de l'Université Grenoble Alpes

Spécialité : Signal, Image, Parole, Télécoms

Arrêté ministériel : 25 mai 2016

École Doctorale EEATS

Grenoble Institut Neurosciences

Neurotechnologies et Dynamiques des Réseaux

Décodage de la parole à partir de l'activité cérébrale à l'aide de méthodes linéaires

Decoding speech from brain activity using linear methods

Gaël Le Godais

Thèse soutenue publiquement le **28 juin 2022** devant le jury composé de:

Frank Guenther
Professeur, Boston University
Rapporteur

Fabien Lotte
Professeur, INRIA Bordeaux Sud-Ouest
Rapporteur

Olivier David
Directeur de Recherche, INS Marseille
Examinateur

Tetiana Aksenova
Directeur de Recherche, CEA Grenoble
Examinatrice

Anne Guérin-Dugué
Professeur des universités, UGA
Présidente du Jury

Blaise Yvert, *Directeur de thèse*
Directeur de Recherche,
INSERM Grenoble

Gaël Le Godais

Décodage de la parole à partir de l'activité cérébrale à l'aide de méthodes linéaires
Decoding speech from brain activity using linear methods
Signal, Image, Parole, Télécoms, 28 juin 2022
Rapporteurs: Frank Guenther et Fabien Lotte
Directeur: Blaise Yvert

Université Grenoble Alpes
Grenoble Institut Neurosciences
Neurotechnologies et Dynamiques des Réseaux
École Doctorale EEATS
Bâtiment Biologie B, 2280 rue de la Piscine
38400, St Martin d'Hères, France

Abstract

Invasive brain-computer interfaces controlled by paralyzed people could restore natural speech production by providing real-time speech synthesis from cortical activity. This thesis aims at decoding existing invasive recordings of speech activity in an offline setting, using real-time compatible methods that could later be used in a natural speech brain-computer interface.

A focus was made on decoding speech from cortical activity using linear methods, in particular partial least squares regression, which has been successfully used in motor brain-computer interfaces before but not for speech decoding yet. Two main approaches were compared: 1. direct decoding of F0 and mel cepstral coefficients of speech, and 2. indirect decoding of speech through an articulatory representation. In order to decode articulatory trajectories from cortical activity, those were first inferred from the patient's audio recordings using dynamic time warping. Several feedforward and recurrent neural networks were trained on a separate electromagnetic articulography dataset to perform articulatory-to-acoustic synthesis, and were evaluated using objective and perceptive criterions. The best model was finetuned to predict mel cepstral coefficients of speech from decoded articulatory trajectories. Speech was synthesized from decoded F0 and mel cepstral coefficients using an MLSA filter, for both decoding paradigms.

Both direct and indirect decoding of acoustic features of speech achieved significant speech decoding with similar performances, although not intelligible. Partial least squares regression was found to perform a more efficient feature reduction than PCA-based linear regressions, for a similar performance. Prior to decoding, noisy channels and spectral features of cortical activity that do not contain speech information were successfully removed using an automatic feature selection. It was found that decoding from spectrograms of cortical activity was best when using all selected frequencies up to 200Hz and concatenating the last 200 ms of brain activity. Decoding of articulatory trajectories was significantly better from frontal electrodes than from temporal electrodes, and the opposite was true for acoustic features of speech. However, in both cases decoding was significantly better when including all electrodes. Finally, our experiments suggest that decoding could be improved by splitting a speech decoder into a voicing classifier and a regression-based decoder only active on voiced segments.

In this thesis, we set up an entire real-time-compatible decoding pipeline based on linear methods. It should now be implemented for further evaluation in a close-loop experiment. Meanwhile, although decoding was much better than chance, linear methods are likely not good enough yet for a brain-computer interface generating natural speech. Further work should focus on developing real-time compatible decoders based on other methods like deep neural networks.

Résumé

Le contrôle d'une interface cerveau-ordinateur invasive par une personne paralysée pourrait restaurer une production naturelle de la parole en permettant une synthèse vocale en temps réel à partir de l'activité corticale. Cette thèse vise à décoder de manière hors ligne des enregistrements invasifs existants de l'activité corticale de la parole, en utilisant des méthodes compatibles temps réel qui pourraient ensuite être utilisées dans une interface cerveau-ordinateur générant de la parole naturelle.

L'accent a été mis sur l'utilisation de méthodes linéaires pour le décodage de la parole à partir de l'activité corticale. En particulier la régression des moindres carrés partiels, qui a déjà été utilisée avec succès dans des interfaces cerveau-ordinateur moteur, mais pas encore pour le décodage de la parole. Deux principales approches ont été comparées : 1. le décodage direct de F0 et des coefficients mel cepstraux de la parole, et 2. le décodage indirect de la parole via une représentation articulatoire. Afin de décoder les trajectoires articulatoires à partir de l'activité corticale, celles-ci ont d'abord été déduites des enregistrements audio du patient à l'aide d'un algorithme de déformation temporelle dynamique. Différents réseaux de neurones récurrents ou à propagation avant ont été entraînés à effectuer une synthèse articulatoire-acoustique sur des données d'articulographie électromagnétique, et ont été évalués à l'aide de critères objectifs et perceptifs. Le meilleur modèle a été ajusté par fine-tuning à prédire les coefficients mel cepstraux de la parole à partir des trajectoires articulatoires décodées. La parole a été synthétisée par un filtre MLSA à partir de F0 et des coefficients mel cepstraux décodés.

Le décodage direct et indirect des caractéristiques acoustiques de la parole ont atteint des performances similaires, toutes deux significativement meilleures que la chance bien que non intelligibles. La régression des moindres carrés partiels s'est avérée effectuer une réduction des caractéristiques neurales plus efficace que les régressions linéaires basées sur l'ACP, pour une performance similaire. Avant d'effectuer le

décodage, les canaux bruyants et les caractéristiques spectrales de l'activité corticale qui ne contiennent pas d'information sur la parole ont été supprimés avec succès par une sélection automatique. Nous avons constaté que le décodage à partir de spectrogrammes de l'activité corticale était optimal lors de l'utilisation de toutes les fréquences sélectionnées, jusqu'à 200Hz, et en concaténant les 200 dernières millisecondes d'activité cérébrale. Les électrodes frontales ont permis de mieux décoder les trajectoires articulatoires que les électrodes temporales, tandis que l'inverse était vrai pour les caractéristiques acoustiques. Dans les deux cas cependant, le décodage a été significativement meilleur en utilisant toutes les électrodes à la fois. Enfin, nos expériences suggèrent que le décodage pourrait être amélioré en divisant le décodeur en un modèle classifiant le voisement d'un côté et une régression active uniquement sur les segments vocaux de l'autre.

Dans cette thèse, nous avons mis en place un pipeline de décodage complet basé sur des méthodes linéaires et compatibles temps réel. Il devrait maintenant être implémenté pour une évaluation plus approfondie dans une expérience en boucle fermée. En parallèle, bien que le décodage soit bien meilleur que la chance, les méthodes linéaires ne sont probablement pas encore assez performantes pour être utilisées dans une interface cerveau-ordinateur produisant de la parole naturelle. De prochains travaux devraient se concentrer sur le développement d'autres décodeurs compatibles temps réel, basés notamment sur des réseaux de neurones.

Acknowledgement

Writing a PhD thesis is no easy task, and going through those years would not have been possible without the support of many.

I cannot thank enough my thesis director Blaise Yvert for providing me support and a good working environment. Your positive energy was always motivating me and your steady direction has been keeping me on the right track. It was a real pleasure to work with you. I would like to thank my reviewers, Professor Frank Guenther and Professor Fabien Lotte, for their valuable feedback and interesting discussions. I would also like to thank Dr. Olivier David and Dr. Tetiana Aksenova for following up with my PhD and always giving me food for thoughts, as well as Professor Anne Guérin-Dugué for kindly accepting to be the president of my jury.

Thanks to my Gipsa friends, the lab was a much nicer place. I will miss your company and remember dearly our adventures. Omar, Remi, I am forever thankful for your support, I could not imagine finishing my PhD without you. Thank you to all the people of my second lab for being so welcoming, we had a good time. Marie and Philémon, I am so glad we did this together, no one could dream of better PhD mates. Florent, I would like to thank you for all your the work on which this PhD is based and for all your help. To the next generation of students, I wish you all good luck, I have no doubt you will do great. Of course I also have to thank all my old friends for their support, you always managed to make me laugh and have fun.

Thank you to my partner Mia for enduring me during writing, it was not always easy for you. You had the patience to deal with my incapability to balance work and personal life and you did a lot to make things easier for me. I am very lucky to have you.

My family played an invaluable role for supporting me through my PhD. Thank you to my parents for supporting me through my studies. Thank you to my brother for standing by our mom when times were rough and I was not there to help. And thank you to my mom, for your endless support and for never giving up. You are one of the strongest persons I know, if it was not for you finishing my PhD would have never been possible.

Finaly, thank you to everyone who is taking time to read this. I hope this work will be useful to you.

This work was supported by the French National Research Agency under Grant Agreement No. ANR-16-CE19-0005-01 (Brainspeak), and by the European Union's Horizon 2020 research and innovation program under Grant Agreements No. 732032 (BrainCom).

Contents

General Context

Motivation and Problem Statement

In France, about 600 patients are reported to suffer from locked-in syndrome (LIS), and 5000 to 7000 patients are diagnosed with amyotrophic lateral sclerosis (ALS). Many of those patients preserve the cognitive ability to speak and communicate while loosing the physical ability to do so. When possible, communication protocols using remnants of voluntary movements can be used to interact with the patients and improve their quality of life, however those methods remain slow and often require assistance.

Recently, brain-computer interfaces (BCIs) have been fueling hope for more efficient communication systems and overall improved independence, with the possibility to also be used by patients that do not retain any motor control. Non-invasive BCIs that do not require surgery typically focus on typing based communication and computer control. Ideally however, a communication BCI would allow for natural speech production. Such speech BCI would require a fine control of multiple degrees of freedom, which is not possible with non invasive recordings. Research on motor BCIs showed that invasive recordings can provide good enough neural signals to control robotic limbs with multiple degrees of freedom. A BCI could possibly be designed to produce natural speech from invasive electrodes.

Considering that the implantation of a subject with invasive electrodes causes medical risks, a natural speech BCI cannot be directly developed with a subject. Its potential should first be validated on recordings of patients that were implanted for medical reasons. This thesis aims at decoding speech from existing invasive recordings in an offline setting, focusing on real-time compatible methods that could later be used in a close-loop speech BCI. The possible designs for a natural speech BCI include either the control of acoustic descriptors of speech, or the control of a set of virtual articulators. The later requires to build an articulatory-to-acoustic speech synthesizer. Finally, patients stay only a few days in the hospital with implanted electrodes. So only a few sessions of recordings are possible, which limits the number of sentences. This is a limitation for training deep neural networks to decode speech. We investigated the extent to which linear methods, which require fewer data, can be

used to successfully decode speech from brain signals. Investigating those technical challenges of speech decoding from cortical activity would get one step closer to prototyping a natural speech BCI.

Thesis Structure

This thesis is composed of the following chapters:

Chapter 1 describes the context and associated state of the art of the thesis. It introduces the core principles of speech production and perception, including a description of the underlying articulatory, acoustic and cortical mechanisms. It describes relevant recording methods and mathematical representations for each mechanisms, with a focus on speech synthesis techniques and cortical models of speech. It presents both invasive and non invasive brain-computer interfaces, and reviews the literature on brain-computer interfaces for communication and speech decoding from invasive recordings.

Chapter 2 details the objectives of this thesis.

Chapter 3 describes the materials and methods of the thesis. It presents datasets of articulatory, acoustic and neural recordings of speech and their respective processing. It details articulatory-to-acoustic speech synthesis methods, speech decoding from neural activity, and their respective evaluations.

Chapter 4 presents results of speech decoding from neural activity using different linear regression and feature reduction methods. It compares direct decoding of acoustic speech representation from neural activity with indirect decoding of speech through an articulatory representation. Different methods of articulatory-to-acoustic synthesis used for indirect decoding are evaluated. Additionally, this chapter explores decoding of a F0 and formants using a conjunction of regression and classification. Finally, it briefly presents preliminary results of speech decoding using neural networks.

Chapter 5 summarizes the results presented in chapter 4 and discusses their implications and limitations.

1 State of the Art

1.1 Principles of speech articulation

Human speech relies on a set of physiological structures that allow for sound production via manipulation of airflow. While some of these structures are passive, some can be finely controlled to produce a wide variety of distinct sounds. The elemental sound units produced by human speech called **phones** are combined into meaningful speech representations like words or sentences. Different languages would use different set of phones and combine them in different ways to produce meaningful speech. However the underlying physical mechanisms of speech production are shared between speakers and therefore are language agnostic. This section describes what are these physiological mechanisms that produce speech sounds.

1.1.1 Anatomy of speech organs

Speech relies on manipulation of airflow to generate specific sounds. The main structures responsible for manipulating the airflow produced by the lungs are represented in Figure 1.1. Air is expelled from the lungs into the larynx, through the vocal folds, and into the pharynx. The airflow is then routed to the nasal cavity and/or to the oral cavity where a set of articulators further shape the incoming airflow.

1.1.1.1. Lungs

The **lungs** are two air reservoirs located in the thoracic cavity. They are connected to the **trachea** which is itself directly connected to the **larynx**. The base of the lungs is attached to the **diaphragm**, a skeletal muscle responsible for most of the airflow control. When contracting, the diaphragm extends the lungs to increase their volume. This creates a lower pressure in the lungs that draws air in. On the contrary, relaxing the diaphragm combined with the contraction of intercostal muscles and the lungs elasticity pushes air out. While these mechanisms are mainly supporting

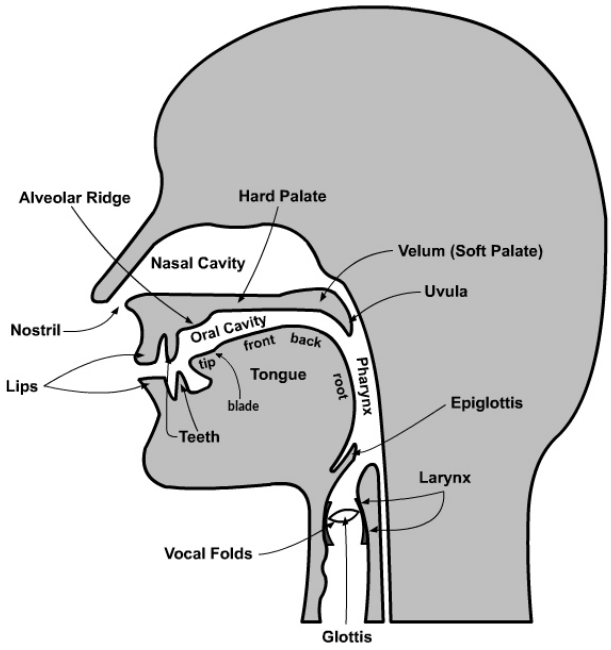


Fig. 1.1.: Midsagittal diagram of speech organs.
adapted from *MadBeppo.com, 2021*

the respiratory system by enabling oxygen extraction from the air, they are also providing the controlled airflow necessary for speech.

1.1.1.2. Larynx

The **larynx** is an organ located in the upper neck, largely made of cartilages as well as tissues and muscles. Aside from protecting the trachea from food, it can constrict to manipulate airflow and most importantly contains the **vocal folds**. The vocal folds are two membranes that are stretched horizontally across the larynx, their opening controlled by a set of muscles. This structure allows them to be pulled together and vibrate under airflow.

A representation of relaxed open vocal folds letting air to flow is shown in Figure 1.2. Bringing the vocal folds together while pushing air from the lungs creates pressure on them. Depending on the airflow and the closure of the vocal folds, multiple phenomenons can appear:

- **Vocal fry**, a creaky voice sound where air is accumulating behind the folds and going through on a periodic basis creating a periodic pop sound.
- **Modal voice**, the combination of glottal pressure and air flow makes the vocal folds smoothly vibrate, creating a pitched sound rich in harmonics.

- A noisy, airy sound when the larynx is constricted but the vocal folds remain open, typically found during whispering.

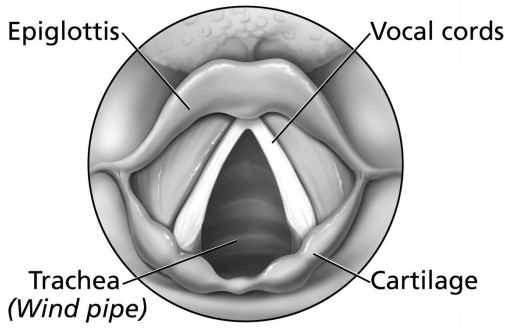


Fig. 1.2.: **Larynx from above.** Air flows along the trachea through the larynx from below. The vocal folds are represented open here with a upside down V shape, the open ends can be brought together to close the vocal folds and vibrate from airflow. from *National Cancer Institute*

1.1.1.3. Hard and soft palates

On the roof of the mouth, in between the **oral** and **nasal cavity** lies the **palate**. The front part is referred to as **hard palate** as it is bony and rigid. The back part is called **soft palate** or **velum**. It is a moveable muscular mucous membrane without bone that controls how much air is sent through the nasal cavity or the oral cavity. It can be lifted to completely close the nasal cavity or lowered to let more air through the nose. The very end of the soft palate ends in a narrower extension called **uvula** susceptible to vibrate.

1.1.1.4. Lips

The **lips** are two opposing soft, fleshy parts of the face closing the oral cavity. The **upper lip** can be lifted and the **bottom lip** can be lowered. The corners at the junction of both lips can be retracted, lowered or raised. The lips can therefore be closed or open to various degrees either vertically (**aperture**) or horizontally. This allows to shape the mouth from completely closed and compressed to wide open, including a semi closed state leaving only a small opening. Lips can also be pushed forward (**protrusion**) or rolled inward (**compression**), rounding the lips further away or closer to the teeth. The soft structure of the lips combined with the set of muscles controlling them also allows them to vibrate with airflow or completely seal the oral cavity.

1.1.1.5. Teeth

Inside the oral cavity lie the **teeth**, hard calcified structure mainly dedicated to eating. Only the bottom teeth are movable through the activation of the **mandible** (lower jaw, see 1.1.1.7), which they are fixed on. While the teeth are mostly passive elements of the oral cavity for speech production, the front teeth in particular interact with the lips and tongue. With some degree of closure, the tongue can constrict or seal the oral cavity against or in between the teeth.

1.1.1.6. Tongue

The **tongue** is a muscular organ with a high degree of mobility provided by its four intrinsic and four extrinsic muscles. The tongue lies both in the pharynx where its root is located and in the oral cavity as shown in figure 1.1. The oral part of the tongue is typically split in three regions: the **tip** (or *apex*), the **front** and the **back** (or *dorsum*). A fourth region right behind the tip, on the top of the tongue is often distinguished from the tip: the **blade**. The tongue does not only belong to the oral cavity as its **root** lies back in the pharynx. The tongue can move in such a way that any of its four oral regions can interact with other parts of the oral cavity such as the teeth, lips, palate and uvula. The fine motor control of the tongue allows for precise dynamic gestures like clicks, drills, constriction or even complete sealing of the oral cavity.

1.1.1.7. Jaw

The jaw is an articulated structure housing the oral cavity used to manipulate food. It is composed of the fixed **upper jaw** and the moveable **lower jaw** also respectively called *maxilla* and *mandible*. The lower jaw controls the opening of the oral cavity by rotating around its fixation axis at the back of the mouth. As it supports the lower teeth, lower lip and tongue, the opening angle of the mandible is implicated in speech production. Any gestures from the lips or tongue needs to be coordinated with it.

1.1.2 Mechanics of speech production

The mechanical production of speech is based on the physical interaction of the speech organs described in section 1.1.1 with air. While the articulators have

intrinsically complex continuous and dynamic motions, each elementary speech sound (**phone**) they produce can be well described by a set of simple articulatory characteristics. In this section we broadly describe the mechanical link between articulation and spoken language by detailing how phones are produced. We will simply name **gesture** any articulatory gesture producing an elementary speech sound. In order to describe the articulatory characteristics of a phone, we will sometimes use minimal pair examples: pair of words with only one phone difference that differ by only one (when possible) articulatory characteristic. For example, *'sock'* and *'shock'* only differ by their first phones written as */s/* and */sh/*.

1.1.2.1. Phonation

A speech sound produced while vibrating the vocal folds is categorized as **voiced**. On the contrary, a sound produced with the vocal folds apart is **voiceless** or **unvoiced**. In English the first consonants of *'vain'*, *'zen'*, *'game'* are voiced whereas *'fame'*, *'sane'*, *'came'* are not. Voiced sounds are most often **modal**, which means the "true" vocal folds are smoothly vibrating, like in the previous examples. However some breathy or creaky voiced vowels can also be found in some languages (Gordon, 1998). It is also not uncommon in American English to hear a creaky voiced counterpart of phones that are normally modal, although those do not carry a separate linguistic information.

In a modal voiced sound, the **pitch** generated by the vocal folds can be controlled to provide linguistic meaning. Combined with rhythm, loudness and timbre, **intonation** - or variation of pitch - constitute **prosody**. Intonation can contribute to stressing syllables, provide phrase boundary tones or add emotional information (Gussenhoven, 2004, Chapter 2). More importantly, pitch height and/or variation can be discriminative word characteristics in tone languages. Register tone languages would distinguish syllables with low or high pitch, with sometimes multiple shades in between. Contour tone languages would distinguish sounds based on a rising, falling, or stable pitch trajectory. Some languages such as Mandarin even distinguish more complex pitch trajectories like a rising then falling tone or a falling then rising tone during a single syllable (Gussenhoven, 2004, Chapter 3).

1.1.2.2. Places of articulation

Obstructing airflow in different places of the vocal tract will produce different sounds. While there is a wide shade of places of articulation, we can classify them in three

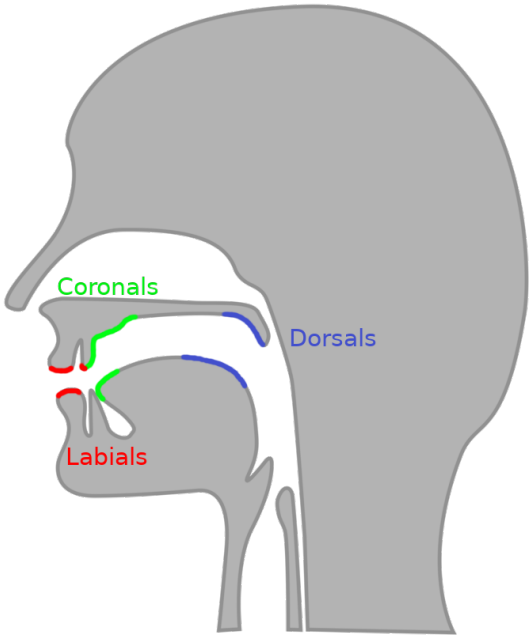


Fig. 1.3.: **Main places of articulation.** Red parts of the articulators interact in labial articulations, green for coronal articulations and blue for dorsal articulations.

main categories: 1. **labial** articulations where the obstruction is performed by the lips, 2. **coronal** articulations where the tongue performs the obstruction in the region of the teeth and alveolar ridge, and 3. **dorsal** articulations using the back of the tongue. The three involved regions are represented in figure 1.3 in a midsagittal diagram of speech organs.

1. Labials: Articulations performed with two lips such as in the consonants from *'pie', 'buy'* or *'my'* are called **bilabials**. Articulations can also be **labiodental**, with the lower lips touching the upper teeth such as in *'fee'* and *'vie'*.

2. Coronals: Common coronal articulations where the tongue tip or blade presses against the upper teeth (and sometimes lower teeth) are called **dental** (*'thigh', 'thy'*). In **alveolar** articulations the tongue tip (or blade) presses against the alveolar ridge (depending on the speaker: *'tie', 'die', 'nigh', 'sigh', 'zeal', 'lie'*). Lastly, articulation places further back in mouth where the alveolar ridge and the palate meet like in *'shy'* or *'she'* are called **post-alveolar**.

3. Dorsals: Sounds produced by pushing the back of the tongue against the soft palate can be found at the end of *'hack', 'hang'* or *'hag'*. Those are **velar** articulations. **Palatal** articulations where the front of the tongue is against the hard palate are also sometimes classified as dorsals, although sometimes also as coronals.

1.1.2.3. Nasality

During most speech, the soft palate is raised enough so that air goes through the oral cavity. When it is lowered enough it prevents the air from going through mouth and directs it through the nose. When saying *'ram', 'ran'* or *'rang'* the air starts by going through the mouth on the *'ra'* and ends up only through the nose during the last consonant. Holding the last consonant of these words and then pinching the nose illustrate it. Sounds produced with a lowered soft palate are called **nasal**. When the soft palate is not completely lowered, air goes both through the mouth and the nose, creating **nasalized** sounds.

1.1.2.4. Manners of articulation

Multiple kinds of articulatory gestures can be performed across the different places of articulation. Those gestures are referred to as **manners of articulation**.

Articulators can perform a complete closure of the oral cavity that prevents all airflow from leaving the mouth, characterizing **stop** consonants. **Plosives** are a common case of stop consonants produced with a lifted soft palate preventing air from going through the nasal cavity (Ladefoged and Johnson, 2015, Part 1.1). As the airstream is completely obstructed by the articulators, pressure builds up in the mouth and is released as a burst. An example of bilabial plosives would be found at the beginning of *'pie'* or *'buy'*, with the lips coming together to block airflow and releasing pressure in a pop sound. The closure can also be performed by the tongue against the roof of the mouth in alveolar plosives such as in *'tie'*, *'dye'*, or velar plosives: *'kye'*, *'guy'*. Consonants described in section 1.1.2.3 also respectively present bilabial, alveolar and velar closure, however letting air through the nose. This reduces the pressure build up in the mouth, removing the sharp burst sound of the consonant.

When two articulators get very close without obstructing the vocal tract, that creates a turbulence in the airflow characteristic of **fricative** consonants. Common English fricatives can be labiodental (*'fie', 'vie'*), dental (*'thigh', 'thy'*), alveolar (*'sigh', 'zoo'*), or palato-alveolar (*'shy'*).

There are many other manners of articulation like trills, taps, or glottal stops. Some manners are a combination of different manners like **affricates** which are stop consonants releasing as fricatives (first consonant of *'cheer'*). Others are in

between consonants and vowels: **approximants** perform constriction of the vocal tract without creating turbulence. That can be achieved by partial closure like the consonant from *'lie'* where the tongue lets air flow on the sides. It can also be achieved by a loose narrowing of the vocal tract that does not produce a fricative turbulence (*'we'*).

1.1.2.5. Consonants

From an articulatory point of view, consonants are formed by obstructing airflow in the vocal tract. Places and manners of articulations described earlier in this subsection are therefore characteristics of consonants. Hence why the examples given so far focus on consonants.
Given the classification of sounds we described in this section, a consonant is characterized by its:

1. Voicing
2. Place of articulation
3. Nasality
4. Manner of articulation

For example, the consonant in *'my'* is a voiced bilabial nasal consonant. The manner of articulation is omitted since its 'nasal' characteristic implies it is a nasal stop consonant. As another contrasting example, *'sigh'* starting consonant is a voiceless alveolar sibilant consonant, a special case of a voiceless coronal fricative. It's nasality is omitted as the soft palate is lifted.

1.1.2.6. Vowels

During vowel production, the articulators do not obstruct the vocal tract like during consonant production. To be precise, the articulators do come closer together, only not near enough to constrict the vocal tract and generate turbulences in the airstream. Vowels are mostly characterized by the position of the tongue and the lips, although voicing and nasality can also distinguish vowels in some languages (for example french nasal vowels).
The tip of the tongue typically rests behind the lower front teeth while the front or back of the tongue arcs upwards, as shown in figure 1.4 by a set of 4 vowels from words *'heed'*, *'hard'*, *'hoard'* and *'heard'*. The height of the tongue will produce

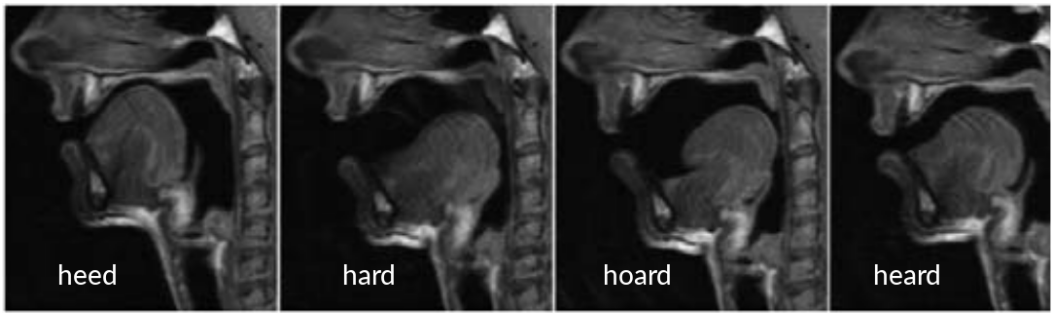


Fig. 1.4.: MRI of New Zealand English vowels in the midsagittal plane. From left to right, the images show articulation of the sustained vowels in heed, hard, hoard and heard.
from *Watson et al., 2009*

distinctive sounds: in *'heed'* the front of the tongue is close to the palate while in *'hard'* it is much lower. However these examples share a frontal position of the tongue, the highest point of the tongue being in the mouth. Other vowels exhibit a tongue pulled toward the back of the vocal track: *'father'* with a low tongue, or *'good'* with a higher tongue.
In combination of the mouth, the lips can either be wide open or come closer to alter the color of the vowels. **Rounding** occurs in *'who'd'* while vowels like in *'heed'* are unrounded.

1.1.3 Recording techniques

With exception of the lips and jaw, most articulators are hidden with little visibility from outside. Two main methods are used to track the movements of the articulators: 1. imaging techniques 2. position tracking based techniques. Due to the highly dynamic and precise nature of speech production, one has to meet high spatial and temporal precision in order to accurately record speech articulation. Some consonants such as plosives (which will be described in) exhibit especially fast dynamics requiring a resolution of ~10ms.

On one side the imaging techniques provide a visual representation of the vocal tract through the skin. Those can directly be read for medical diagnosis but also analyzed to extract the positions of the articulators on the video recording. Common imaging techniques include **X-ray imaging**, **ultrasonography** and **magnetic resonance imaging** (**MRI**, example in figure 1.4). Out of these, only MRI provides clear vision of the articulators. Recent MRI improvements allow for real-time imaging at a resolution of 33ms while keeping good spatial resolution of 1.5mm (Niebergall

et al., 2013). While this does not perfectly meet requirements of time and spatial resolution, it is a viable and safe method for real-time recording of speech articulatory trajectories. Its main drawback remains the high cost of the MRI scanner and the difficulty to access such a system.

On the other side, the position tracking-based techniques provide cheap high time resolution recordings of speech articulators. They include **electromyography** (**EMG**), **electropalatography**, and **electromagnetic articulography** (**EMA**). Electromyography records the electrical activity of speech muscles by sticking electrode arrays on the face. However this does not directly provide the positions of the articulators and their reconstruction is a difficult problem. Electropalatography uses electrodes mounted on an artificial palate molded to the subject. The electrodes measure the contact points of the tongue against the palate during speech, however no other information about articulators is recorded. Finally, electromagnetic articulography stands out by its good resolution and coverage of the articulators.

1.1.3.1. Electromagnetic Articulography (EMA)

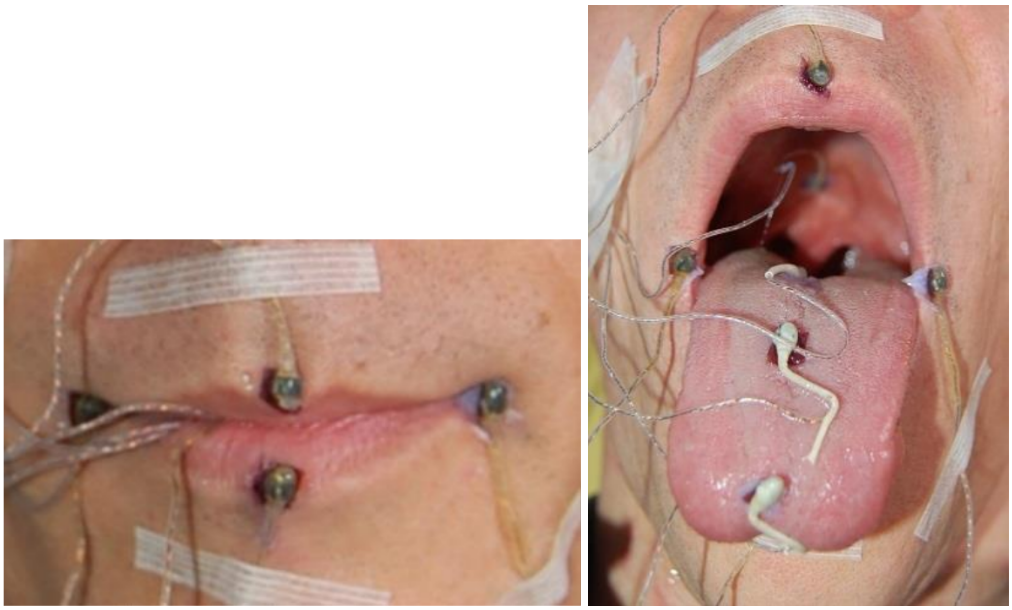


Fig. 1.5.: **Electromagnetic articulography**. EMA coils are glued to the lips, the jaw, the tongue and the soft palate.
from *Bocquelet et al., 2016c*

Electromagnetic articulography provides 3D recordings of a limited number of selected points of the vocal tract but with high spatial and temporal resolutions of about 0.1mm and up to 500Hz (Schönle et al., 1987; Zierdt et al., 1999).

A variable electromagnetic field is produced by a set of induction coils mounted next to the subject's head, with each induction coil producing its own frequency. Sensor coils are glued on several points of the subject's articulators and plugged into an acquisition system. When the subject speaks, each sensor coil moves accordingly to its articulator. This movement in the electromagnetic field induces an electrical current in the sensor coils which is recorded by the acquisition system. As the intensity of the electromagnetic field is proportional to the inverse cube of the distance from its transmitter, the voltage induced in the sensors is directly related to their distance from the induction coils. Since each induction coil generates an electromagnetic field with a different frequency, it is possible to reconstruct the distance from the sensor to each induction coil, thus locating the sensor in space by triangulation.

In order to record meaningful points of the vocal tract, the positions of the sensor coils have to be chosen carefully. The most important parameters to measure are arguably the lower jaw, the lips protrusion and opening, the tongue constriction points, the velum and vocal folds. The design of EMA does not allow to record either the pharynx constriction or the vocal folds activity. Due to the rigidity of the jaw, its opening can accurately be recorded with a single coil, for example on the lower teeth. The lips can be recorded by gluing sensors on their corner and on the middle of the lower and upper lip. As we explained in the previous sections, the tongue creates constrictions with its tip, front and back, which are all good candidates to glue the sensors. Finally the velum opening can be well approximated with a single sensor.

Considering that most information about articulation belongs to the midsagittal plane, EMA recordings can be performed in 2D as well. However 3D recordings are more robust to head tilt, which can be compensated using a sensor glued to a fixed part of the head.

EMA recordings suffer from a few limitations:

1. It only allows to record a few select points of the vocal tract, although this can be mitigated by choosing carefully which points to record. Unfortunately it is not always possible to record the velum as some subjects are especially sensitive and cannot tolerate this sensor. The sensors and their cables going out of the mouth are indeed a bit uncomfortable. More importantly they slightly impair speech by disturbing the normal articulatory trajectories. The recorded speech is therefore not natural, although perfectly intelligible.

2. EMA recordings have to be done in one sitting. Once a sensor is detached, it is impossible to find again the exact location where it was glued. This limits the amount of coherent data one subject can record. Moreover the recording session can be interrupted if a coil detaches during the recording because of salivation. Typical experiment duration is limited to a maximum of 1 to 2 hours before a coil detaches.

3. The recording setup is not very invasive and not very expensive but it requires a bit of preparation and space. It can be difficult or even impossible to setup with other recording techniques, especially because of the electromagnetic field.

Despite these limitations, EMA remains a reliable real-time recording method that provides direct access to the articulator's trajectories. This is especially useful for mathematical analysis of speech and computer applications. In particular, EMA recordings have been successfully used for articulatory to speech synthesis from offline recordings (Toda et al., 2008; Liu et al., 2016) and in a real time closed loop setting (Bocquelet et al., 2016c).

1.2 Speech Acoustics

In a spoken conversation, a speaker encodes language into sound through the control of articulators, and a listener decodes language from perceived sounds. To satisfy a proper communication, both speech sounds and articulatory gestures share the same language information. This section first presents the acoustic properties of speech sounds embedding linguistic information and their corresponding articulatory gestures. Then, common acoustic representations of speech are presented as well as speech synthesis techniques.

1.2.1 Phonation

Figure 1.6 shows a transcription of the English sentence *They do not understand that* spoken by an American male native speaker (extracted from a TV interview: `https://youtu.be/JsZqHHb8BnE?t=9`). The spectrogram shows the evolution of the frequency content of the speech along the sentence. Most of the spectrogram during speech shows a clear stripe pattern, evidence of the modal phonation. The vibration of the vocal folds generates a sound with high harmonic content. Its fundamental frequency **F0** ranges typically from 100Hz to 150Hz for men and

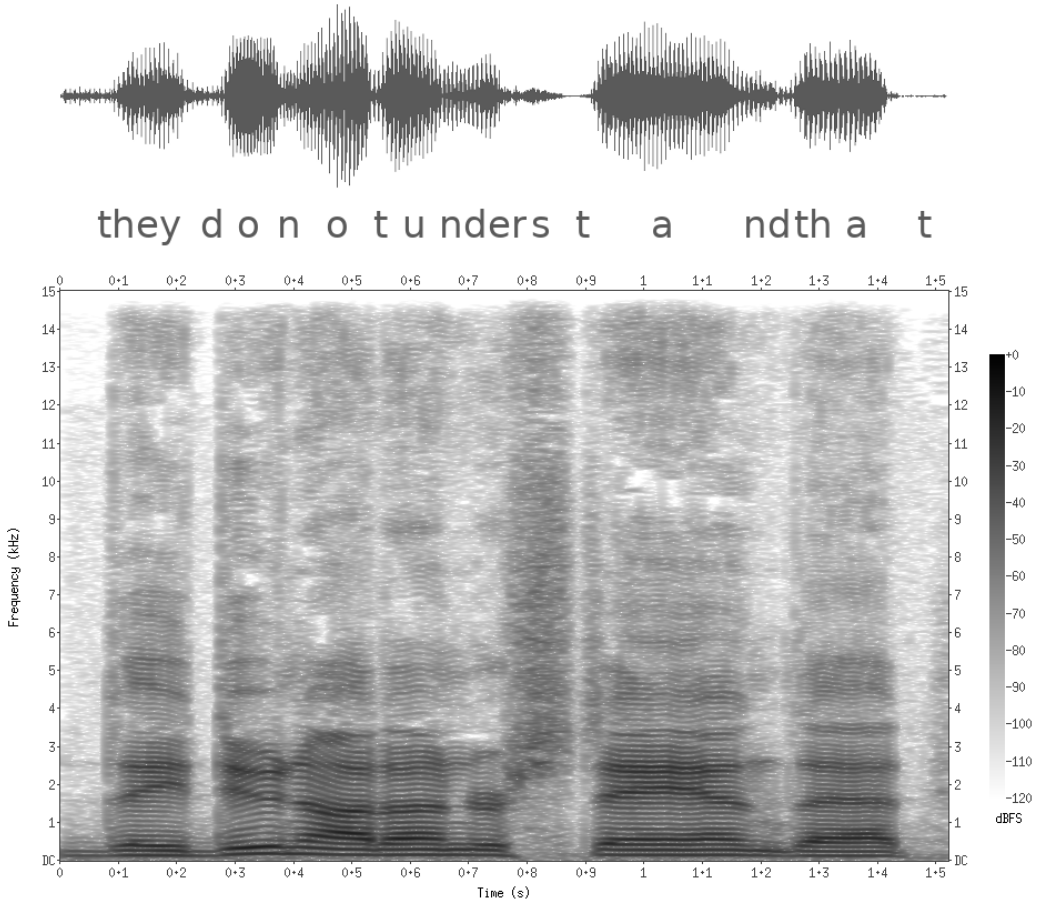


Fig. 1.6.: Spectrogram and waveform of sentence *'They do not understand that.'* spoken by a male American speaker.
The transcription of the sentence was synchronized with the waveform and spectrogram to exhibit the relationship between acoustics of speech and language. Frequencies are represented on the vertical axis of the spectrogram from 0 to 15 kHz and time on the horizontal axis from 0 to 1.5 seconds. Power of a given frequency was represented on a grey scale from white (no power) to black (maximum power).

180Hz to 250Hz for women (Eriksson, 1995). Depending on intonation, F0 can rise and fall outside of these average values. The dark stripes in the spectrogram of figure 1.6 represent harmonics *ie* of F0, which cover all the integer multiples of F0 with varying amplitudes. On the contrary, voiceless parts of the sentence can be recognized on the spectrogram by the absence of these stripe patterns. It can clearly be seen on the *'s'* sound of *understand* between 0.8 and 0.87 seconds, where the spectrogram shows darker areas without harmonic content indicating noise.

1.2.2 Acoustics of Vowels

Many experienced blowing into a glass bottle to produce a defined note. Although the air blown into the bottle is merely a noise, the natural resonance of the bottle focuses it into a narrow frequency band, creating a perceivable pitch. Filling the bottle with water will reduce its cavity volume, thus lifting the bottle's resonant frequency. This can actually be easily heard when filling a water pitcher from the tap, the pitch produced by the water getting higher and higher as the water pitcher fills up. Now this phenomenon called Helmholtz resonance is very close from what happens in the vocal tract when producing vowels.

The vocal tract during vowel production can be approximated as a long tube open at one end (the lips) and closed at the other (the glottis). When the tongue lifts to produce vowels, it creates a succession of cavities and necks with characteristic resonant frequencies. Those resonant frequencies are called **formants**, usually noted from lowest to highest **F1**, **F2**, **F3** and **F4** although the two first formants are usually enough to distinguish vowel sounds (Bladon and Fant, 1978). Vowel formants have been measured for the first time by Chiba and Kajiyama (Chiba and Kajiyama, 1941) by computing the vowels spectra and looking at their peaks (a reproduction from Arai, 2004 is shown in figure 1.7). Formants can also be distinguished in vowel spectrograms like in figure 1.6: for example during the */a/* sound at $1.35s$, dark lines indicate higher amplitudes at the approximate frequencies F1=600Hz, F2=1600Hz, F3=2500Hz and F4=3600Hz.

Plotting the vowels on a diagram with axes F1 and F2, a quadrilateral organization appears (see figure 1.8). In normal speaking conditions, all vowels first formants will fall into this quadrilateral as it is constrained by the acoustic properties of the vocal tract. Our previous example is no exception to this, falling near $/\varepsilon/$. It is also possible to produce a very similar diagram by plotting the vowels on a plane with axes being the tongue height and front/back position (fig. 1.8). Such planes are referred to as spatial-planes since they classify vowels based on the spatial position of

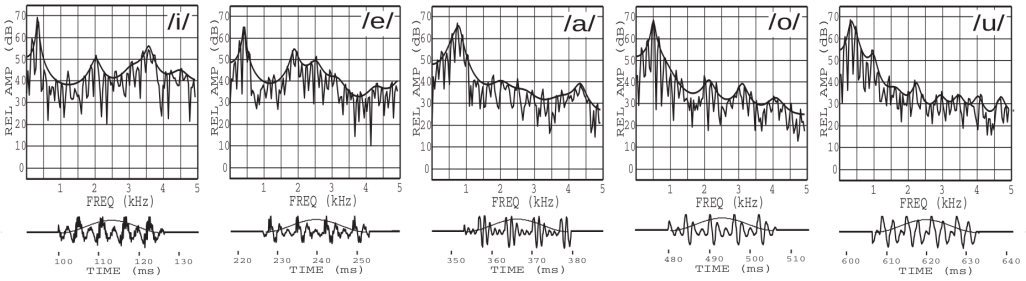


Fig. 1.7.: Reproduction of vowels spectra in Chiba and Kajiyama. The Fourier analysis of vowel sounds highlights spectral peaks (formants) characterizing the different vowels.
from *(Arai, 2004)*

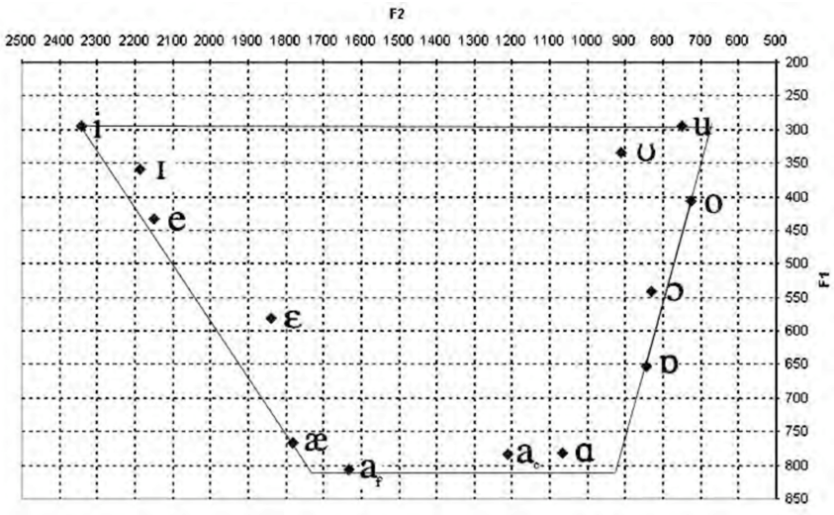
the tongue. In order to classify vowels using a secondary articulation (like nasality) it is possible to define a secondary plane dedicated to those vowels.

1.2.3 Acoustics of Consonants

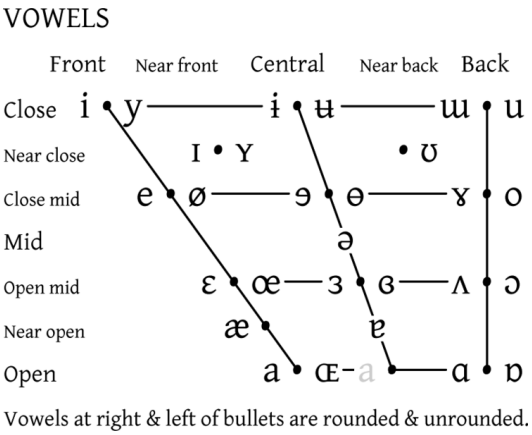
By definition and contrasting with voiceless consonants, voiced consonants are supported by a modal voice. As described in subsection 1.2.1, the consonant therefore contains (at least partly) F0 as well as all its harmonics. It can also exhibit formants and/or influence formant transitions at the vowels onset and offset.

We defined plosives as consonants containing a complete closure of the vocal tract followed by a burst release. The complete closure implies a moment of silence, while the burst creates a short transitory period. In figure 1.6, this is very clear at $0.9s$, where both the spectrogram and the waveform show a moment of silence followed by a short noise burst. The */d/* around $0.25s$ is very similar although the vocal folds keep vibrating even with a closed mouth as it is a voiced stop. Through the closed mouth, only the low frequencies can be heard, hence why only the F0 and its first harmonics appear. During the burst, a noisy transitory period occurs before the onset of the following vowel.

Fricatives sounds present noises with different spectral shapes and relative amplitude, depending on their phonation and place of articulation (Jongman et al., 2000). A sibilant */s/* is a strong noisy voiceless sound with a lot of high frequency content, as clearly seen in figure 1.6 at $0.82s$, whereas */th/* consonants are noisy but voiced sounds with much lower pitch and amplitude. The very first consonant at $0.08s$ of figure 1.6 exhibits F0 and its first harmonics while all the upper spectrum is noisy and inharmonic (no stripes above 1kHz). The spectrogram also shows F1 and F2



(a) F1 F2 diagram



(b) IPA vowel chart

Fig. 1.8.: Vowel diagrams (a) F1 F2 diagram of some IPA vowels highlighting the acoustic quadrilateral. Lowest values of F1 and F2 are found on top right, maximum values in bottom left. The possible combinations of F1 and F2 are roughly delimited by 4 straight lines: the acoustic quadrilateral. It is highlighted by a set of vowels that exhibit maximal and minimal F2 values (from *Hitch, 2017*). (b) IPA Vowel Chart. Vowels are organized in a diagram relatively to the tongue's position. The tongue's front/back position is represented on the horizontal axis and the degree of closure it performs is represented on the vertical axis (from *International Phonetic Association, 2018*).

with low intensities during */th/* and characteristic transitions with the following vowel.

During nasal consonants such as */m/*, or */n/* the mouth is closed while air goes through the nose. The mouth becomes essentially a close side branch of the nasal cavity, having complex interactions with the resonant frequencies. The oral cavity absorbs the energy of frequencies close to its natural resonant frequencies, which creates anti-resonances in the output sound (Fujimura, 1962): **anti-formants**. Not only these anti-formants actively substract some frequencies from the output spectrum, but it also reduces all higher formants amplitude. Having sound resonating in both the oral and nasal cavities has another consequence on the sound color: as the total volume of air and the total surface of the cavity walls increase, they absorb more energy from the sound, which results in a sound with dampened resonance. This results in formants with broader bandwidth compared to non nasal sounds.

1.2.4 Source-filter model

In the previous sections 1.2.2 and 1.2.3 we showed how the shape of the vocal tract controlled by the articulators modulates a sound source to produce a wide variety of sounds, where the sound source can often be the glottis like in the case of voiced sounds, or a constriction in the vocal tract like in the sibilant */s/*. This understanding of speech acoustics is called **source-filter** theory of speech production (Fant, 1981).

A simple model of the vocal tract as Helmholtz resonators has already been used in section 1.2.2 to explain formants. This model has been around long before the modern understanding of source-filter theory of speech production (Müller, 1839; Helmholtz, 2009). Despite its simplicity it does allow for decent prediction of formants (Maekawa, 2002). However, proof that vowel production is determined by vocal tract shape was settled by Chiba and Kajiyama (Chiba and Kajiyama, 1941). They created mechanical models of the vocal tract during vowel production based on actual measurements of the vocal cavity. In order to simplify the models they made it as straight tubes with varying section, using a pulsating telephone receiver diaphragm to model the vocal folds. Precise replications of these models (Takayuki, 2001) can be seen in figure 1.9. Out of the 5 Japanese vowels produced by the mechanical models, 4 were rated highly intelligible following subjective listening tests.

As technology allowed to encode and decode sounds using electrical circuits, it became possible to use signal processing to model the speech source and vocal

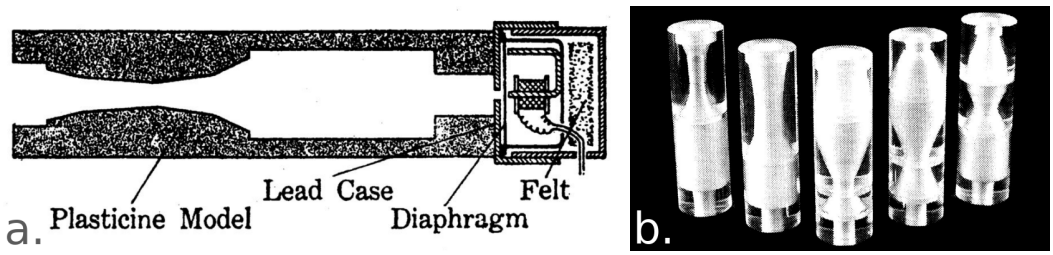


Fig. 1.9.: Mechanical models of vowels.
a. Chiba and Kajiyama's mechanical model for vowel */i/* (Chiba and Kajiyama, 1941, p.129)
b. Precise reproductions for 5 Japanese vowels (from left: */i/*, */e/*, */a/*, */o/*, */w/*) from Arai (Takayuki, 2001)

tract as electrical signals and filters. This permitted a finer analysis and modelling of speech sounds, source and filters (Fant, 1981). Computers and digital signal processing pushed it further, using the same methods on a discrete representation of sound. The applications of source-filter theory of speech production include efficient and robust sound representation dedicated to speech as well as speech synthesis.

1.2.5 Speech representations

1.2.5.1. Fourier analysis

The Fourier analysis of a time signal provides a powerful way to study its frequency components, a detailed explanation can be found in appendix A.1.1.1. The resulting **spectrum** can be useful to study short stationary segments of speech such as vowels. However the inherent dynamic nature of speech cannot be properly analyzed with a spectrum, as the spectral components of two consecutive phones can be vastly different and would be average in a spectrum. For this reason, **spectrograms** are commonly used to study speech's spectral components over time. Technical details are described in appendix section A.1.1.2.

Although both methods are useful tools for analysis of time signals, they are not optimized for speech. Multiple methods designed for computing efficient speech representations are presented next.

1.2.5.2. LPC

Linear Predictive Coding or **LPC** is a class of methods using linear predictive models to perform a compressed representation of digital speech. The original LPC

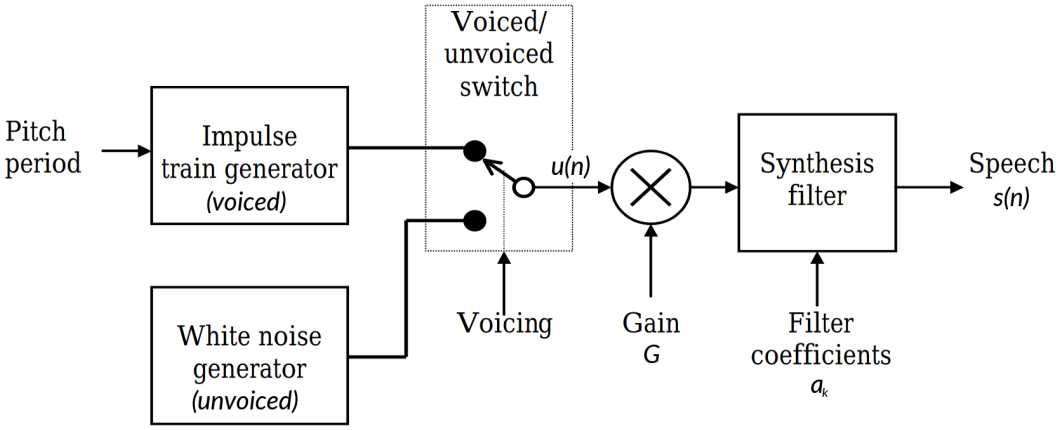


Fig. 1.10.: The LPC model of speech production. A source signal $u(n)$ is generated from an impulse train for voiced sounds and a white noise for unvoiced sounds. The speech signal $s(n)$ is computed by applying the filter coefficients a_k to the source signal.
adapted from *Foundation and Evolution of Standardized Coders* (Chu, 2003)

standard is based on the source-filter model of speech (Tremain, 1982; Chu, 2003) by encoding the voiced or voiceless source on one side, and the spectral contributions of the glottal flow, articulators and vocal tract on the other side.

Under the assumption that speech has stable characteristics over a short span of time, the speech signal is processed into short nonoverlapping frames. The voicing, eventual pitch period, overall gain and filter coefficients modelling the response of the vocal tract are computed on each frames independently. The source is modelled either by a random noise for voiceless sounds or a periodic impulse train for voiced sounds. The resulting LPC model of speech can be visualized in figure 1.10.

The LPC filter design uses the assumption that for a given speech signal $s(n)$, its n^{th} sample can be approximated by a linear combination of its p previous samples $n-p, n-p+1, \ldots n-1$. Considering a source $u(n)$, a gain G and filter coefficients a_k, a speech signal $s(n)$ is modelled by LPC as:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{1.1}$$

LPC provides a very efficient and robust compression of speech, however the many assumptions and approximations are responsible for deterioration of the overall quality. Over the decades, many improvements have been brought to the very simple design described in this paragraph while conserving the principle of source-filter model of speech (Chu, 2003).

1.2.5.3. Cepstrum

A powerful propriety of the Fourier transform is given by its **convolution theorem** stating that a convolution in the time domain is transformed into a multiplication in the frequency domain:

$$\mathcal{F}\{f * g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\} \tag{1.2}$$

For a speech signal, the source and the filter are convoluted in the time domain and multiplied in the frequency domain. The spectrum of a voiced sound contains $F0$ and all its harmonics, which amplitudes are multiplied by the formant resonances of the vocal tract. By applying a logarithmic transformation to the spectrum, the source and filter are now additive in the **log spectrum**.

The **power cepstrum** (or **magnitude cepstrum**) is defined as the inverse Fourier transform of the log power spectrum. Given a time signal $x : \mathbb{R} \to \mathbb{C}$, its power cepstrum C is defined as (Bogert, 1963) :

$$C(\tau) \triangleq \left| \mathcal{F}^{-1} \left\{ \log \left(\left| \mathcal{F} \left\{ x(t) \right\} \right|^2 \right) \right\} \right|^2 \tag{1.3}$$

Let a speech sound signal $x : \mathbb{R} \to \mathbb{C}$ such as $x(t) = s(t) * h(t)$, with $s : \mathbb{R} \to \mathbb{C}$ the source signal and $h : \mathbb{R} \to \mathbb{C}$ the filter signal. Let their Fourier transforms $X(f), S(f)$ and $H(f)$. The power cepstrum of $x(t)$ is equivalent to the power spectrum of the sum of the log power spectrums of $s(t)$ and $h(t)$, effectively decomposing the source and filter signals:

$$\begin{aligned}
C(\tau) &= \left| \mathcal{F}^{-1} \left\{ \log \left(\left| \mathcal{F} \left\{ s(t) * h(t) \right\} \right|^2 \right) \right\} \right|^2 \\
&= \left| \mathcal{F}^{-1} \left\{ \log \left(|S(f)|^2 \cdot |H(f)|^2 \right) \right\} \right|^2 \\
&= \left| \mathcal{F}^{-1} \left\{ \log \left(|S(f)|^2 \right) + \log \left(|H(f)|^2 \right) \right\} \right|^2 \\
&= \left| \mathcal{F}^{-1} \left\{ \log \left(|S(f)|^2 \right) \right\} + \mathcal{F}^{-1} \left\{ \log \left(|H(f)|^2 \right) \right\} \right|^2
\end{aligned}$$

As the vocal folds source and the formant filters have vastly different frequency dynamics, they contribute to vastly different **quefrencies** (equivalent of frequencies in the cepstral domain). As the spectral peaks of the voice's F0 and its harmonics are by definition spaced by a frequency gap equal to F0, the voiced source of speech has a quefrency peak around its $F0$. This makes the cepstrum a valuable tool to extract the $F0$ of speech (Noll, 1967) . Unfortunately the formants do not exhibit a clear representation in the quefrency domain. Another drawback of the power cepstrum is the loss of the phase information, which is useful for speech synthesis.

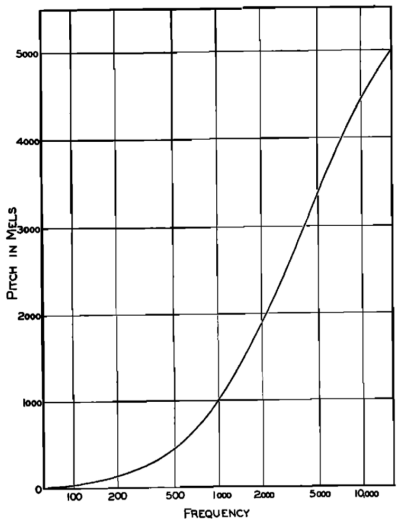


Fig. 1.11.: Original mel scale. The figure displays the subjective relation between pitch and frequency as measured by Stevens et al., 1937

As an alternative, the **complex cepstrum** removes the computation of the spectral power of the Fourier transforms in order to keep the imaginary part of the Fourier transform, which contains the phase information. It is defined as:

$$C_c(\tau) \triangleq \mathcal{F}^{-1}\left\{\log\left(\mathcal{F}\left\{x(t)\right\}\right)\right\} \tag{1.4}$$

1.2.5.4. Mel-frequency Cepstrum

Human perception of pitch does not scale linearly with frequency: it has a higher resolution at low frequencies, meaning that the difference between a 100 Hz and a 150 Hz pure tones feels much bigger than between a 3000 Hz and a 3050 Hz pure tones. The **mel scale** has been defined as a non linear frequency scale matching a linear perception of pitch (Stevens et al., 1937). It is an experimental scale measured by subjective perception tests and defined so that a 1000 Hz pure tone has a perceptual pitch of 1000 mels. Many approximations have been proposed to match the mel scale by a mathematical function (Umesh et al., 1999), but there is no exact theoretical formula to it.

The **Mel-frequency Cepstrum** provides speech features based on the power cepstrum that includes information about human pitch perception. Let $x : \mathbb{R} \to \mathbb{C}$ a time signal, $P : \mathbb{R} \to \mathbb{R}$ its power spectrum computed by DFT such that:

$$P(f) = \left|\mathcal{F}\left\{x(t)\right\}\right|^2 \tag{1.5}$$

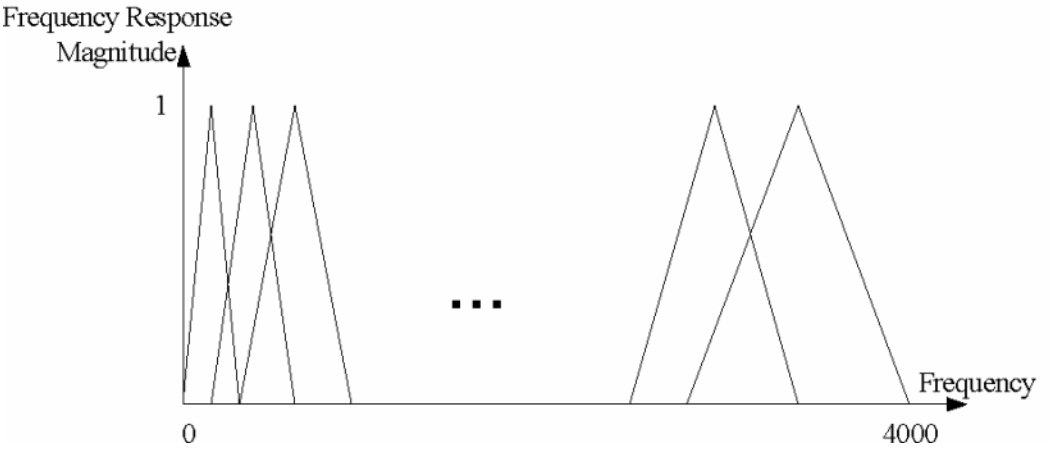


Fig. 1.12.: Mel-frequency filter bank. A set of K overlapping triangular filters covering the whole frequency spectrum of a 8kHz signal. The delimiting frequencies of each triangular filter are equally spaced on the mel scale.
from *Han et al., 2006*

The power spectrum P is integrated on equally spaced frequency bands on the mel scale using filter banks. The classic implementation uses triangular filter banks, $K = 35$ filters achieving good performance (Zheng et al., 2001). Given a set of $K + 2$ equally spaced frequencies $(f_k)_{0 \leq k \leq K+1}$ with $f_0 = 0$Hz and $f_{K+1} = F_s/2$ the Nyquist frequency, we can define a set of triangular filters $(T_k)_{1 \leq k \leq K}$:

$$T_k(f) \triangleq \begin{cases} 0 & \text{if } f < f_{k-1} \text{ or } f > f_{k+1} \\ \dfrac{f - f_{k-1}}{f_k - f_{k-1}} & \text{if } f_{k-1} \leq f < f_k \\ -\dfrac{f - f_{k-1}}{f_k - f_{k-1}} & \text{if } f_k \leq f < f_{k+1} \end{cases} \tag{1.6}$$

A visual representation of such triangular filters can be seen in figure 1.12. Let $I(k)_{1 \leq k \leq K}$ the coefficients resulting from the integration of the power spectrum over the filter bank:

$$I(k) = \sum_{f=0}^{F_s/2} P(f) T_k(f) \tag{1.7}$$

The mel-frequency cepstrum is computed from $\log(I(k))$ by applying a **Discrete Cosine Transform** (**DCT**). The amplitudes of which are called **Mel-Frequency Cepstral Coefficients** (or **MFCCs**):

$$MFCC(d) = \sum_{k=1}^{K} \ln(I(k)) \cos\left[d(k - 0.5)\frac{\pi}{K}\right], \quad d = 0, \ldots, D \tag{1.8}$$

Due to the proprieties of the DCT, most information lies into the first coefficients of the DCT. It is common practice to only compute the first $D \leq K$ coefficients as a lossy compression, typically $D = 12$. MFCCs encode a lot of relevant speech information in only a few coefficients, making them very popular for speech application. However their efficiency is based on tackling only real signals, therefore discarding phase information about the original signal. That unfortunately makes MFCCs unsuitable for speech synthesis.

1.2.5.5. Mel Cepstrum

The **mel cepstrum** is an alternative to MFCCs based on the complex cepstrum (equation 1.4). It is computed by warping the complex spectrum onto a mel frequency scale approximation before computing its logarithm. By using the complex spectrum, the mel cepstrum distinguishes from MFCCs by not computing the power spectrums.

In order to detail its properties we need to briefly introduce the **Z transform**. Given a real signal $x(n)_{n\geq 0}$, we define its (unilateral) Z transform $H(z)$ by:

$$H(z) = \mathcal{Z}\{x(n)\} = \sum_{n=0}^{\infty} x(n) z^{-n}, \quad z = Ae^{j\omega} \in \mathbb{C} \tag{1.9}$$

The Z transform is closely related to the DFT as we can notice by picking $z = e^{2\pi j f}$.

Assuming $c(m)_{0\leq m\leq M}$ the coefficients of the complex cepstrum (eq. 1.4) of a speech signal, the vocal tract transfer function $H(z)_{z=e^{j\omega}}$ is represented by:

$$H(z) = \exp\left(\sum_{m=0}^{M} c(m) z^{-m}\right) \tag{1.10}$$

Similarly, Imai, 1983 showed that $H(z)$ can be represented by the M$^{\text{th}}$ order mel cepstral coefficients $\tilde{c}_\alpha(m)_{0\leq m\leq M}$ wrapped on the mel frequency scale by:

$$H(z) = \exp \sum_{m=0}^{M} \tilde{c}_\alpha(m) \tilde{z}^{-m} \tag{1.11}$$

Where, $\tilde{z}^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}$, and α is a parameter depending on the sampling rate that has been computed by Imai, 1983 to approximate the mel scale. (Imai, 1983)

Like MFCCs, selecting only the first mel cepstral coefficients allows for approximate reconstruction of the voice spectrum. The generally considered optimal order of

the mel cepstrum is $M = 24$ (25 coefficients). Although it does not provide a representation of speech quite as compressed as MFCCs, the mel cepstrum avoids computing the power spectrums, therefore retaining phase information. That feature makes it a good speech representation candidate for speech synthesis applications.

1.2.6 Speech synthesizers

Speech synthesis has been used for various applications ranging from text-to-speech synthesis to efficient voice compression for telephone networks and music applications.

Text-to-speech synthesis is commonly performed by **concatenative synthesis**: elementary speech units are concatenated using linguistic knowledge extracted from text. Units are recorded speech samples compiled into a database of sounds. Concatenative synthesis usually provides the best speech synthesis quality. However it is speaker dependant by design and requires linguistic knowledge.

Telephone networks historically required heavy voice compression, relying on efficient representation of speech. Instead of a linguistic representation of speech, **vocoders** analyze acoustic characteristics of sound, typically spectral envelope and source signals. Vocoders are therefore speaker and language independent, to the cost of a lower performance compared to concatenative synthesis.

In this section we will focus on vocoders, discussing some commonly used systems based on a mel cepstral representation of speech.

1.2.6.1. MLSA

A **Mel Log Spectrum Approximation** (**MLSA**) filter is designed to synthesize speech sound waves from a mel cepstrum and an excitation signal (Imai, 1983). The filter approximates the speech spectral envelope represented by the mel cepstral coefficients, where its excitation signal models the speech source.

In order to use the MLSA filter, one needs to compute the mel cepstral coefficients $\tilde{c}_\alpha(m)$ of a speech signal described in equation 1.11 . Possible methods include using linear predictive coefficients (Tokuda et al., 1994) or adaptive algorithm (Fukada et al., 1992).

On the other hand, one needs to extract the voicing from the speech signal. As we saw the cepstral representations are very well suited to detect the fundamental

frequency of speech. However several other efficient methods have been designed to extract F0 from speech signals, notably RAPT algorithm (Talkin, 1995), SWIPE (Camacho and Harris, 2008) and REAPER (*REAPER* 2021). Once the voice source is extracted, the excitation signal of the MLSA filter can be modelled: it can either be a white noise for voiceless sounds or an impulse train with a period $1/F0$ for voiced sounds.

During the synthesis, short overlapping consecutive frames of sound are computed separately by using windowing. The mel cepstral coefficients computed from each frame typically introduce discontinuities between each frames. Using interpolation, the MLSA filter introduces little spectral distortion, making them suitable for robust high quality speech synthesis. Moreover the MLSA can be implemented for near real time synthesis.

Although MLSA filter is the only method used to synthesize speech in this manuscript, two recent alternative methods are described next.

1.2.6.2. WORLD

Vocoder-based synthesis is often computationally expensive, especially so with higher quality techniques. The **WORLD** vocoder combines three analysis and one synthesis algorithm that provides high quality synthesis fast enough for real-time synthesis (Morise et al., 2016).

WORLD also uses the compact F0 and mel cepstrum as speech representation and adds an aperiodicity parameter. Voiced sound sources are indeed never perfectly periodic, and some speech sounds contain a mixture of periodic and aperiodic sources, like voiced fricatives. The binary modeling of speech source using either white noise or impulse train cannot tackle these sounds. The aperiodicity parameters model the non periodic part of the source signal that is not captured by the binary model of noise + impulse train. The three features of WORLD analysis are combined into its synthesis module which generates sound waves. The reported quality of WORLD vocoder is higher than other high quality vocoders while achieving much faster computation time, effectively achieving real-time without reducing sound quality (Morise and Watanabe, 2018). Although the synthesis has very high quality, it does not yet achieve seamless reconstruction of original speech, leaving room for improvement.

1.2.6.3. Neural network based vocoders

In recent improvements on text-to-speech synthesis, convolutional neural networks directly predicting sound signals sample by sample achieved state of the art performance (Oord et al., 2016). Following the success of **Wavenet**, the original neural network performing such a task, improvements were made to make it real-time. By using acoustic features like F0 and mel cepstral coefficients as input instead of text, such models can be used as real-time vocoders. **FFTNet**, a real-time implementation inspired by wavenet, reports higher quality speech synthesis than MLSA, although speaker-dependant (Jin et al., 2018). In order to achieve real-time performance, the small size of the FFTNet can only be trained to represent one speaker. Other neural based vocoders are being investigated with promising results (Valin and Skoglund, 2019; Tian et al., 2020), targeting real-time and natural speaker-independent synthesis.

1.3 Cortical basis of speech production

Speech stems from complex audio and articulatory coding of language. As speech originates from cognitive processes, those should account for both 1. the analysis of speech sounds into a meaningful language representation and 2. the translation of language into motor control of the articulators. This section presents current cognitive models of speech, the associated brain structures and how to record them.

1.3.1 Physiology

The brain is the central organ of the human nervous system, integrating sensory information and mostly responsible for controlling the other organs of the body. It consists of the **cerebrum** on the top, the **cerebellum** below it, and the **brainstem** that connects to the spinal cord. The cerebrum consists of two communicating hemispheres of similar architecture. The surface (**cortex**) of each hemisphere is folded, the rifts (**sulci**) delimiting regions called **gyri** (Fig. 1.13). Those regions can be regrouped into 4 main lobes: the **frontal**, the **parietal**, the **temporal**, and the **occipital** lobe. While interconnected, each lobe hosts specialized and sometime lateralized processes that can be attributed to specific areas.

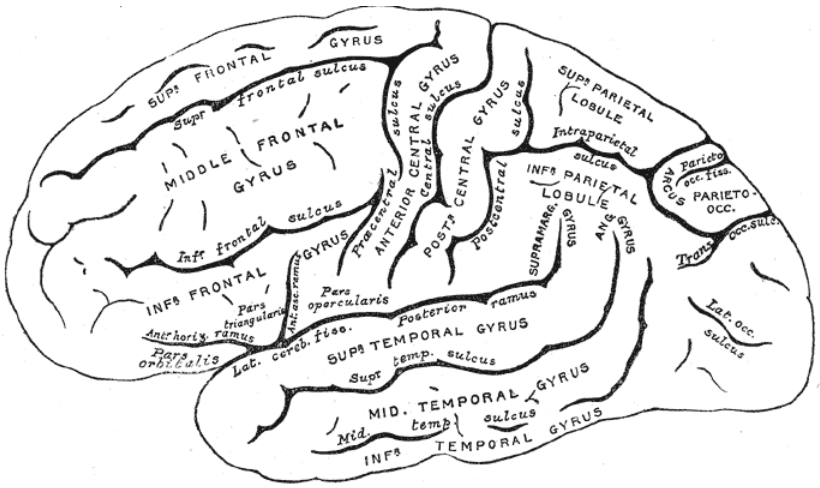


Fig. 1.13.: Major structures of the cerebrum. Representation of the left hemisphere of the cerebral cortex with the frontal lobe on the left and occipital lobe (back) on the right.
from *Gray, 1878*

The brain processing power stems from the conjunction of the **glial cells** and the **neurons**. The neurons are cells forming interconnection with each other and communicating through chemical messengers called **neurotransmitters**. A neuron typically receives signals at the **dendrites** from other neurons' **axons** through connections called **synapses** (Fig. 1.14). Neurons maintain a difference of electrical potential across their membrane by actively exchanging charged ions through the cellular membrane. Signals received at the synapses can be excitatory or inhibitory by increasing or decreasing the neuron's potential. If the potential increases over a short period of time and crosses a threshold, the neuron generates an **action potential** (or **spike**) along its axon which will send a signal to the connected neurons. The propagation of signals through synapses and the propagation of action potentials through the axons rely on a circulation of ions in and out of the cells, which induces **local field potentials** in the extracellular space surrounding the neurons.

1.3.2 Recording methods

Brain dissections have been providing early insight on cognitive mechanisms (see introduction of section 1.3) by studying brain structures and lesions. In the last century, multiple recordings techniques have been developed to study functioning brains (Schultz et al., 2017). Those techniques monitor either indirect markers of neural activity like hemodynamic, or electrophysiological dynamics.

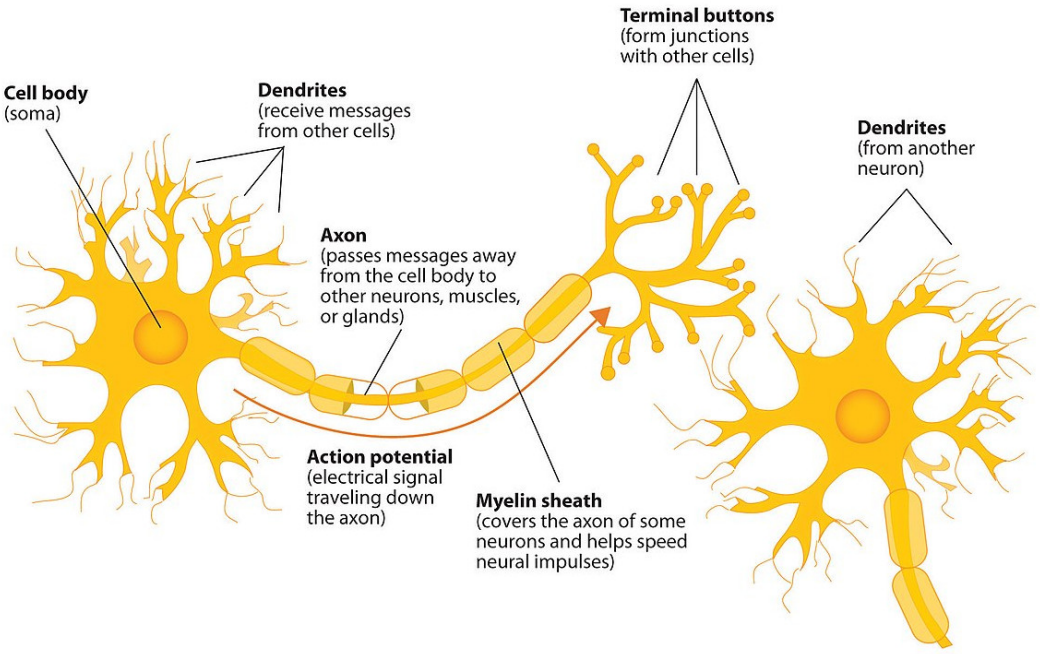


Fig. 1.14.: **Basic neuron diagram.** A simple neuron is represented with a connection from its axon to another neuron. Action potentials run from the soma to the terminal buttons through the axon, a signal is then sent through the synapse to the dendrites of the connected neuron.
from *Walinga and Stangor, 2014*

1.3.2.1. Functional Magnetic Resonance Imaging (fMRI)

Functional Magnetic Resonance Imaging monitors blood flow changes using a strong magnetic field. Blood flow typically increases in brain areas where neuronal activity is more important (Logothetis et al., 2001). fMRI detects the changes in oxygen concentration of the blood (Ogawa et al., 1990) allowing to map increased and decreased activity in the brain. The nature of the process allows a good spatial resolution of a few millimeters but a slow temporal resolution: the observed phenomenon of changes of oxygenation is slow and cannot account for fast changes of cortical activity. As a downside, nomad used is prevented by the size of the chamber and noisy environment, while combination with another simultaneous recording is prevented by the strong electromagnetic field.

1.3.2.2. Functional Near Infrared Spectroscopy (fNIRS)

Functional Near Infrared Spectroscopy also exploits blood oxygen level as a marker of neural activity (Jobsis, 1977) using a different method. Skin, tissues and bones are

mostly transparent to near-infrared light (700-900nm) while hemoglobin absorbs it at different rates depending on its oxygenation. fNIRS is not invasive, portable, cheap and easy to setup (Lühmann et al., 2015). However its spatial resolution is more limited than fMRI and although the technique's temporal resolution of 0.01s is significantly higher than fMRI, the observed phenomenon is as slow as for fMRI's and cannot account for fast changes of cortical activity either.

1.3.2.3. Positron Emission Tomography (PET):

Positron Emission Tomography tracks metabolic processes such as glucose consumption as another indirect marker of neural activity (Bailey et al., 2005). A radioactive tracer analogue to glucose is injected to the patient. The gamma rays emitted by the tracer are then detected using a CT-Scan. This allows a spatial resolution of about 1mm and a temporal resolution of about 0.2s (Castermans et al., 2013). PET is non invasive but expose patients to ionizing radiations, it should therefore be used only sporadically. Moreover it has a high operating cost and is not portable.

1.3.2.4. Magneto Encephalography (MEG):

Magneto Encephalography records magnetic fields produced outside the head by electrical currents occurring in the brain (Cohen, 1968). The small electrical currents resulting from neural activity are detected by very sensitive and costly magnetometers. MEG achieves a few millimeters of spatial resolution and about 1000 Hz of temporal resolution as the LFPs are observed from afar (despite the technically higher resolution). MEG is non invasive and does not expose patients to ionizing radiations. However as it requires to be used in a shielded room, it is not portable, as well as being costly.

1.3.2.5. Electroencephalography (EEG):

Electroencephalography is a measurement of the brain's electrical activity using electrodes placed on the scalp (Berger, 1929). Each electrode records activity of the whole brain through the skin, skull and tissues. Combined with a sensitivity to noise from movement and other environmental artefacts (Goncharova et al., 2003), EEG requires complex processing to cancel interferences and reconstruct sources from the recorded signal (Nunez, Srinivasan, et al., 2006). Despite a low signal-to-noise

ratio and spatial resolution, EEG has a high temporal resolution of about 1000 to 2000 Hz, a non invasive set up only involving gluing electrodes on the scalp, and can be done anywhere for a cheap cost. This explains why it is the most common technique used in communication BCI (Wolpaw et al., 2002).

1.3.2.6. Stereo-Electroencephalography (SEEG):

Stereo-Electroencephalography is a minimally invasive technique used in epilepsy monitoring. Between 5 and 18 electrodes are spread along a needle of about 800 μm diameter with wires running inside. Needles are inserted deep in the brain through small holes in the skull. Electrodes record activity at various depth up to several centimeters in the brain but covering less points on the surface. Not having to open the skull lowers risks of the procedure compared to ECoG but still is rather dangerous. For this reason it is reserved to difficult epilepsy cases (Cossu et al., 2005).

1.3.2.7. Electrocorticography (ECoG):

Electrocorticography is an invasive technique measuring the brain electrical activity at the cortical surface. It is usually used to monitor and plan surgery of severe epilepsy (Crone, 1998). The electrodes can remain implanted up to two weeks during which patients can consent to participate in scientific experiments. An ECoG grid can have up to 256 contacts spaced by 3 to 10 mm. Compared to EEG the intracranial recordings are less sensitive to environmental artifacts and do not suffer from spatial blurring due to skull and scalp (Gevins et al., 1994). ECoG also provides very spatially localized recordings with a good spectral bandwidth that includes high-gamma band (70-200Hz).

The work presented in this manuscript exclusively uses ECoG recordings.

1.3.2.8. Microelectrode Arrays (MEA):

Microelectrode Arrays are very small devices recording extracellular potentials of the nearest neurons with great temporal resolution. For example Utah arrays (Maynard et al., 1997) has a spatial and temporal resolution down to a single neuron action potential. This precision comes with a drawback: only a small area of a few square millimeters can be recorded simultaneously by an MEA. The procedure to implant MEAs into the cortex is invasive and rare in humans.

1.3.3 Models of cortical speech networks

The first hypothesis linking brain processes and speech were formulated following the study of aphasic patients (patients presenting a speech disorder due to brain lesion). In 1861, Paul Broca was the first to perform an autopsy on an aphasic patient that progressively lost the ability to speak while retaining speech comprehension or other cognitive functions. From the study of 12 patients with a similar condition, Broca discovered a link between lesions in a specific region of the left frontal lobe and such expressive aphasia. He conjectured that this region, nowadays baptised **Broca's area**, was responsible for speech production.

Shortly after in 1874, Carl Wernicke performed a similar study on patients suffering from receptive aphasia. Those patients retain the ability to speak fluently, however language comprehension is impaired and language production largely meaningless. The analysis of the brain lesions indicated another brain area responsible for speech comprehension: **Wernicke's area**. It is located in the posterior part of the temporal lobe.

While brain imaging techniques confirmed the importance of those brain areas for speech, more recent clinical studies challenged the view of a clear spatial segmentation of speech production and perception in the brain. Recent models promote a more distributed system of speech interfacing many specialized areas of the brain.

1.3.3.1. Hickok-Poeppel Model

The **Hickok-Poeppel model** or **dual-stream model** of speech processing suggests that perceived speech is simultaneously processed into two streams of information: a ventral stream and a dorsal stream (Hickok and Poeppel, 2004; Hickok and Poeppel, 2007). While speech perception was initially believed to be processed by the auditory cortex, it is inconsistent with modern clinical observations of aphasic patients. In particular, the evidence shows that the ability to perceive sub-lexical speech sounds does not predict comprehension deficits: some patients cannot perform syllable discrimination/identification tasks while showing normal word comprehension. Hickok and Poeppel proposed the dual-stream model as a new model of speech processing that solves those inconsistencies, with additional support from fMRI data.

The **ventral stream** processes speech signals to extract semantic representation. It is to some extent a bilateral process happening in parallel in both hemispheres. The **dorsal stream** is involved into translating acoustic speech signals into articulatory

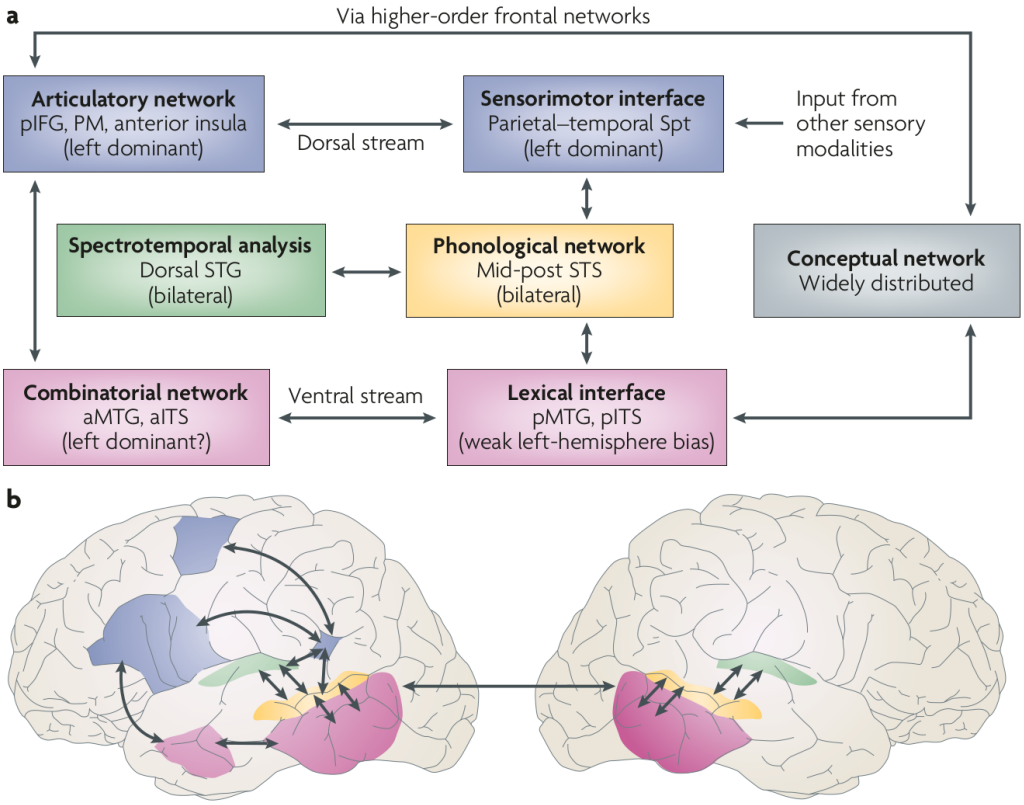


Fig. 1.15.: The dual-stream model of the functional anatomy of language. After initial spectrotemporal and phonological processing of speech, the system diverges into a dorsal stream (blue) and a bilateral ventral stream (pink).
STS, superior temporal sulcus; STG, superior temporal gyrus; aITS, anterior inferior temporal sulcus; aMTG, anterior middle temporal gyrus; pIFG, posterior inferior frontal gyrus; PM, premotor cortex; Spt, Sylvian fissure at the parietal temporal boundary
from *Hickok and Poeppel, 2007*

representations in the front lobe. Contrary to the ventral stream, it is a largely left-dominant process. Both streams are not equally involved in every speech processing task, it is believed that speech recognition tasks engaging lexical access rely more on the ventral stream, while sublexical speech perception tasks rely more on the dorsal stream.

1.3.3.2. DIVA Model

The **DIVA** (**Direction Into Velocities of Articulators**) model describes a neural model of speech acquisition and production accounting for the interactions between motor, somatosensory, and auditory cortical areas (Guenther et al., 1998; Guenther et al., 2006; Tourville and Guenther, 2011). DIVA integrates a feedforward

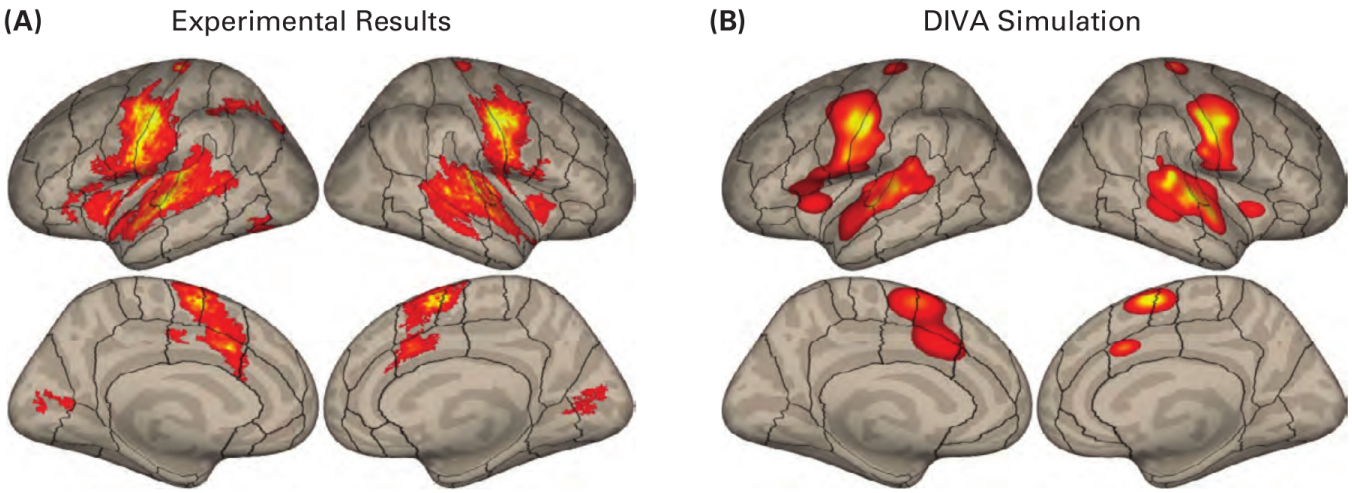


Fig. 1.16.: DIVA modeling of brain activity during syllable production. (A) Brain activity measured by fMRI in subjects reading syllables aloud compared to a baseline of passively viewing syllables. (B) Simulated fMRI activations during syllable production based on DIVA model
from *Guenther, 2016, Chapter 3*

and a feedback subsystem that learn to control a simulated vocal tract by motor commands.

The **feedforward control system** generates known motor commands for articulation of speech sounds while a monitoring system combining auditory and somatosensory feedback checks for the validity of the output and sends feedback motor commands. The **auditory feedback control subsystem** is responsible for detecting and correcting differences between the desired speech sound and the actual auditory feedback. The **somatosensory feedback control subsystem** compares the proprioceptive feedback perception of the articulators with their desired trajectories.

Those subsystems are decomposed into smaller components corresponding to brain regions involved in speech production. A diagram of the detail structure of DIVA model and its neural correlates is shown in Fig. 1.17. Simulated fMRI activity predicted from DIVA model were compared to actual recordings of subjects in Fig. 1.16.

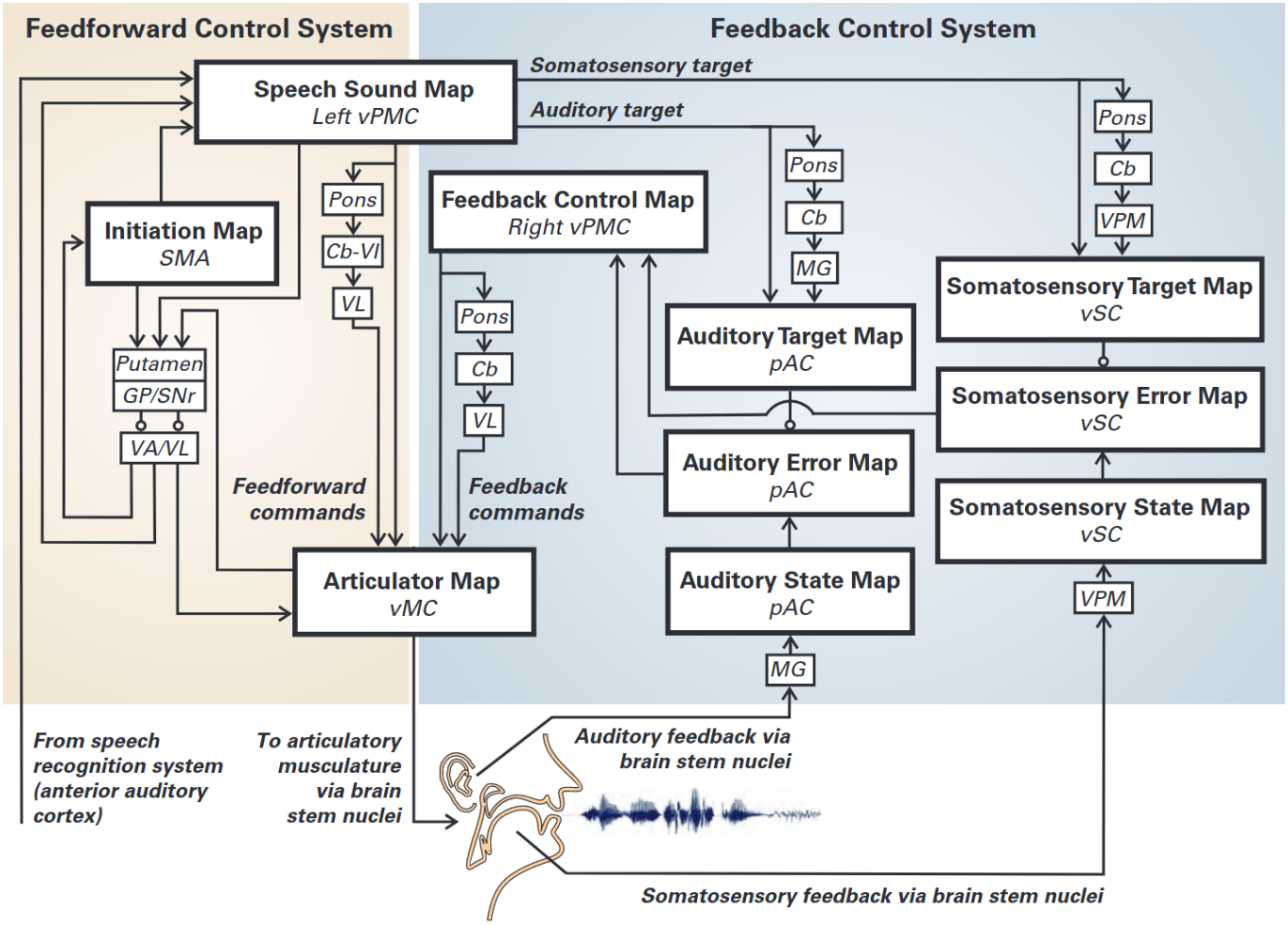


Fig. 1.17.: **DIVA model of speech acquisition and production.** Cb, cerebellum; Cb-VI, cerebellum lobule VI; GP, globus pallidus; MG, medial geniculate nucleus of the thalamus; pAC, posterior auditory cortex; SMA, supplementary motor area; SNr, substantia nigra pars reticula; VA, ventral anterior nucleus of the thalamus; VL, ventral lateral nucleus of the thalamus; vMC, ventral motor cortex; VPM, ventral posterior medial nucleus of the thalamus; vPMC, ventral premotor cortex; vSC, ventral somatosensory cortex
from *Guenther, 2016, Chapter 3*

1.3.3.3. COSMO

COSMO (**Communicating about Objects using Sensory-Motor Operations**) is a cognitive model of speech communication between two agents, a speaker and a listener, aiming to assess theories of the emergence of speech (Moulin-Frier et al., 2015).

In order to model emergence of speech in a multi-agent system, COSMO assumes that 1. each agent possesses a set of prelinguistic abilities; 2. agents must select communication stimuli that are both easy to produce by the speaker and to process by the listener (**adequacy**); 3. the speaker's motor repertoire and the listener's perceptual repertoire must correspond well (**parity**); 4. the speaker and listener must know the correspondence between these motor/perceptual repertoires and the external objects (**reference**).

Following the assumption 1, the agents possess a **motor** and a **sensory** processing system. The conjunction of articulators controlling the vocal tract and cognitive process to control them is denoted M, while the hearing system and its related brain processes are denoted S. The object a speaker wants to communicate is denoted O_S, and the object inferred by the listener O_L. For the communication to be a success, O_S and O_L should be the same, which is denoted by the condition $C \triangleq (O_S == O_L)$.

A resulting model of communication between two agents is shown on figure 1.18. The path from O_S to M represents the speaker's production task, and the path from S to O_L represents the listener's perception task. The production and transmission of sound from the articulators is represented by the path from M to S.

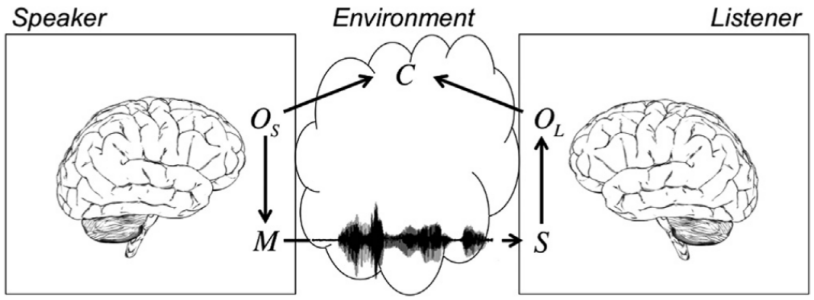


Fig. 1.18.: Schema of speech communication in COSMO. The communication is a success if $C = (O_L == O_S) = 1$
from *Moulin-Frier et al., 2015*

A communicating agent can both be a speaker and listener. Therefore it contains both motor and sensory systems, while having learnt some of the articulatory to acoustic transformation. From this observation, COSMO draws its **central hypothesis**: a

communicating agent can internalize the communication situation (fig. 1.18) inside an internal model (fig. 1.19).

Under this assumption, the internalization model includes: 1. a motor system associating communication objects O_S with motor gestures M; 2. an auditory system associating communication object O_L with auditory stimuli S; 3. a sensory-motor link providing an internal articulatory-to-acoustic model associating motor gestures M with auditory stimuli S; 4. a fusion system assessing whether both internal objects representation O_S from the motor system and O_L from the auditory system are the same.

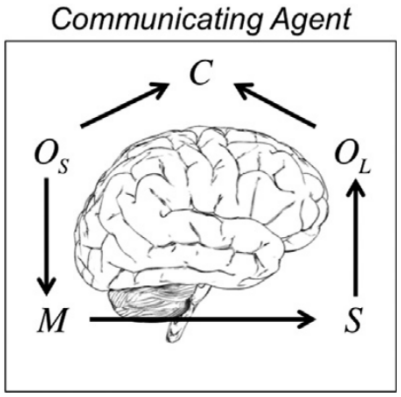


Fig. 1.19.: COSMO's communicating agent internal model of speech.
from *Moulin-Frier et al., 2015*

The internalization hypothesis provides an integrated perceptuo-motor system able to perform both speech production and perception tasks. COSMO acts as an unifying framework of sensory-motor theories (including DIVA), and auditory/motor theories of speech by reversing or deleting some links.

1.4 Brain Computer Interfaces

1.4.1 General Principle

1.4.1.1. Definition

Signals recorded from the brain are giving some information about the actual brain activity. Since the creation The creation of EEG by Berger in 1929, these signals have been interpreted to gain some insight on brain mechanisms, or perform diagnosis. It turns out that those signals can also be willingly modulated. With the advent of

computers, the relative control over the recorded signals opened up the possibility to use brain activity to control a computer program, first described by Vidal, 1973. The brain signals should be processed to extract some control parameters that can be used by an application to control an external device. The interfacing system between brain activity and a computer device is commonly called a **brain-computer interface** (**BCI**) or sometimes a **brain-machine interface**. By extension, BCIs usually describe both the interfacing system and the application that uses it as the distinction is not clear.

The **central nervous system** (**CNS**) consists of the brain and spinal cord. It typically responds to sensory input and produces neuromuscular or hormonal outputs to communicate with the rest of the body and the outside world. While usual tools are controlled by a physical action of some sort, BCIs are opening new possibilities of control solely driven by CNS activity. This fuels hope for restoring motor movements or communication abilities for paralyzed patients, for example by controlling wheelchairs, computers, or even exoskeletons. Moreover, patients that lost the physical ability to speak could benefit from brain driven communication systems.

The following definition of Brain-Computer Interfaces was therefore proposed by Wolpaw et al., 2002: "*A BCI is a system that measures CNS activity and converts it into artificial output that replaces, restores, enhances, supplements, or improves natural CNS output and thereby changes the ongoing interactions between the CNS and its external or internal environment.*"

1.4.1.2. Practical implementation

BCI systems typically consist in supervised machine learning models that attempts to decode the user's neural activity to perform a given task. In order to train the decoding models, BCIs need to be calibrated on a set of data where the expected outcome of the BCI is known. That can be done by running calibration sessions before using the BCI, or by using adaptive learning methods during testing sessions.

Raw brain activity is rarely used for BCI systems. Instead, low dimensional features are extracted from neural activity: spectral information in restricted frequency bands, or characteristic responses to known stimuli. Those features can be used to train classification or regression models. That results into BCIs designed around **discrete** tasks like selection between multiple items, or **continuous** tasks like moving a cursor on a screen.

A BCI system that would not provide real-time feedback to the user runs in an **open loop** setting. On the contrary, in a **closed loop** setting, the BCI provides a real-time feedback to the user. It becomes possible for the user to learn to control the BCI system to perform the expected task. In case of an adaptive training of the BCI system, the BCI and the user mutually adapt to best collaborate.

Finally, the choice of recording methods has a large impact on BCIs. **Non invasive** techniques, like EEG, typically have a limited bandwidth and allow for simple BCI only. **Invasive** techniques however are implanted intracranially and provide higher quality signals, while adding the health risks of a surgical procedure.

1.4.1.3. Patients

First diagnosed in 1966, the **locked-in syndrome** (**LIS**) current definition describes a quadriplegia and anarthria with preservation of consciousness (Smith and Delargy, 2005). Patients suffering from the classic form of locked-in syndrome are typically almost entirely paralyzed, only retaining vertical eye movements. In incomplete forms, patients sometimes retain some remnants of voluntary movements, while other patients exhibit total inability to move and communicate. LIS can stem from multiple afflictions; most commonly brainstem strokes but also traumatic brain injuries, tumors or even infections. Its prevalence is difficult to assess as it is very rare and difficult to diagnose. A study reports a prevalence of 0.7/10000 in dutch nursing homes (Kohnen et al., 2013), while the french locked-in syndrome association (ALIS) currently reports 596 french patients, hinting at a prevalence of around 1/100000 in the general population. LIS is distinct from coma or vegetative states by the preserved consciousness and cognitive abilities of the patients. When possible, vertical eye movements or other remnants of voluntary movements can be used to assess the patients consciousness following communication protocols.

Famously, the late Jean-Dominique Bauby - sufferer of LIS and founder of ALIS - told his experience in the book *Le Scaphandre et le Papillon* only using his left eyelid. While his helper Claude Mendibil would read through the letters *E S A R I N T U L O M D P C F B V H G J Q Z Y X K W*, he would blink his eye to select each character of the book one by one. The letters are essentially the french alphabet sorted from most occurring to the least occurring letter. Even with this optimization, the method remains very slow and impractical.

Other patients experience impairments of communication while preserving cognitive abilities, notably people with **Amyotrophic lateral sclerosis** (**ALS**) also referred as **Lou Gehrig's disease** or **maladie de Charcot** in french. ALS is a neurodegenerative

disease of the human motor system leading 50% of the patients to die within 30 months of the symptoms onset when they do not choose to get intubated, and about 20% of the patients to survive 5 to 10 years (Kiernan et al., 2011). Its incidence rate is reported to be about 2.1 to 4.3 per 100 000 persons-years (Traynor et al., 1999; Mehta et al., 2016). As a consequence of progressive loss of muscle control, dysarthria occurs in about 80% of the cases, effectively reducing the quality of life of the patients (Tomik and Guiloff, 2010). While 20-50% of the patients present some sort of cognitive deficits, many preserve the cognitive ability to interact.

Much like with LIS patients, communication protocols can be designed to maintain interaction with dysarthric ALS patients. Aside from protocols needing a human helper, modern technology allows to control software with remnants physical movements restoring independence of patients. A head and eye tracker allowed former guitar virtuoso and long time sufferer of ALS Jason Becker to write music on a computer (Becker, 2021). Late Stephen Hawking (1942-2018) developed a brilliant academic career as a physicist while suffering from ALS. The famous speech device he used was originally controlled by hand and later on by one of his cheek muscles.

However those devices have limited efficiency as the degrees of freedom of the controls are very low. Moreover they should adapt to the evolution of the patient's physical abilities. Of course, in the case of total locked-in syndrome, no such device or communication protocol based on residual movements can be used. For those patients, diagnosis already relies on interpreting brain activity, consequently brain-computer interfaces raise hope for allowing communication without any physical input. For other patients with limited physical abilities, brain-computer interfaces could improve independence and communication over existing physical control based devices.

1.4.2 Non invasive BCIs

Non invasive methods for recording of brain activity have been used to communicate with ALS and locked-in patients. fMRI has notably been used to assess awareness in patients by asking them yes-or-no simple question. The patients were asked to perform distinct motor and spatial imagery tasks to signify their answer (Monti et al., 2010). This task is useful for diagnosis but fMRI cannot be used for a daily BCI as seen in section 1.3.2.1. In order to facilitate communication with patients, more portable and less expensive methods are favored such as EEG. While EEG based BCI can also be used to answer yes-or-no question, they can also allow free expression

from the user. This section presents common neural features extracted from EEG recordings and their use to control communication BCIs.

1.4.2.1. P300

EEG recorded signals are the combination of noise and the contribution of a large number of small neural activities. Specific stimulus have been shown to induce small albeit reproducible patterns of activity (Sutton et al., 1965). Because of the noisy quality of EEG recordings and the low amplitude of those **event related brain potentials** (**ERPs**), ERPs are hidden in the overall EEG signal. Assuming they are a combination of a Gaussian noise and a reproducible potential pattern time-locked to the triggering event, averaging multiple trials reduces the noise components, effectively improving the signal to noise ratio. The resulting ERPs usually show positive (P) and negative (N) voltage deflections called **components**, which occur at different characteristic times after the stimulus. Among those, the P300 is a positive potential event occurring 250 to 500ms after the stimulus. It is provoked by rare but expected events in a focused task, with a less often occurring event inducing a larger voltage amplitude (Polich, 2007).

In experimental settings P300 can be elicited by expected visual or auditory stimuli. In the **oddball paradigm**, the subject is presented with a series of two stimuli she has to classify. A standard and a target stimulus randomly alternate such that the target occurs rarely, it is the oddball event. The unexpected occurrences of the target stimulus will elicit a P300 ERP in the EEG recording (Donchin, 1981). Alternatively the standard stimulus can be removed altogether, as the random timing of occurrence of the target stimulus is enough to induce P300 of similar quality (Sutton et al., 1967).

P300 BCIs usually make use of the oddball paradigm with either visual stimuli like random flashes (Farwell and Donchin, 1988) or auditory stimuli like tones of different pitch (Käthner et al., 2013). For example, a BCI that would allow to select a cell within a grid can elicit a P300 by flashing in random order the rows and columns. If the subject focuses on a specific cell, the P300 would only occur when its row or column is flashed. By detecting it with a classification method, the cell can be selected. Such a system is suitable for a speller BCI (Fig. 1.20), where each cell contains a character that can be selected to type words (Farwell and Donchin, 1988). P300 spellers have been successfully used to communicate by people with ALS or locked-in syndrome (Sellers et al., 2014). As P300 BCIs are based on attention, they can even be used by patients without gaze control. On the downside, such systems

are quite slow due to the nature of the P300. Although recent improvements pushed the performance of asynchronous P300 spellers up to one character every 3 seconds for a trained user (Townsend and Platsko, 2016), this requires a high cognitive load and patients usually do not reach such performance.

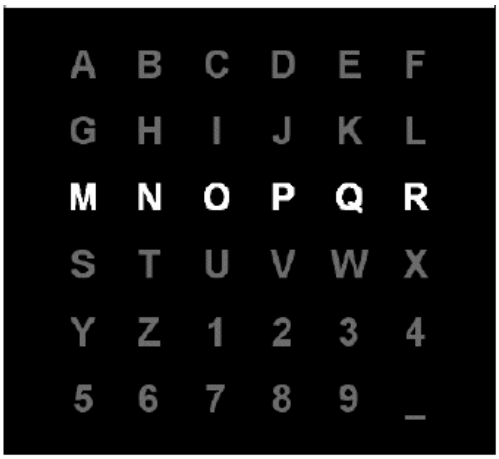


Fig. 1.20.: **Display of a P300 Speller.** One row or column is flashed at a time while the subject focuses on a single character. In this particular example, the line starting with M is currently intensified.
from *Schimpf and Liu, 2013*

1.4.2.2. Steady state visual evoked potential

While P300 is considered an **endogenous** ERP resulting from information-processing activity, **exogenous** ERPs typically occur within 150ms after an external stimuli and reflect activity from the primary sensory systems. Exogenous ERPs induced by sudden visual stimuli are called **visual evoked potentials** (**VEP**). They include N70 and P100 occurring around 70ms and 100ms after the triggering stimuli. A fast periodic flashing visual stimuli like a strobe light triggers periodic sequences of similar patterns, producing a steady-state oscillation. **Steady state visual evoked potentials** (**SSVEPs**) can be analyzed by averaging like P300 but also by frequency analysis. The spectrum of a SSVEP contains peaks at the oscillation's fundamental frequency and its harmonics. Therefore, the frequency of the visual stimuli the subject is focusing on can be reconstructed from the EEG recordings. Stimulus frequencies eliciting SSVEP response used by BCI applications start from 5Hz to more comfortable higher frequency stimulus up to 60Hz (Krolak-Salmon et al., 2003; Wang et al., 2006; Tsoneva et al., 2021).

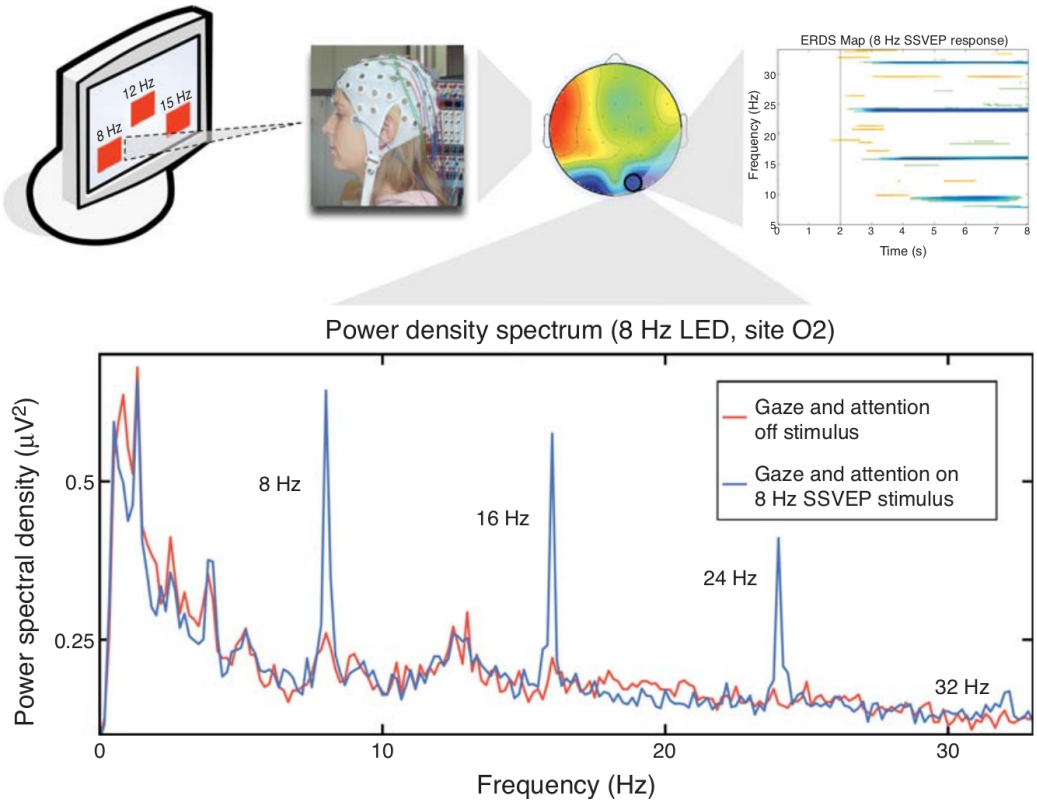


Fig. 1.21.: SSVEP-based BCI operation and analysis. When the user focuses on a red box flickering at 8Hz, a 8Hz EEG activity appears, clearly visible in the power spectrum of the selected occipital electrode.
from *Wolpaw and Wolpaw, 2012*

In the standard SSVEP BCI paradigm, the subject is presented with multiple spread apart repetitive visual stimuli, each of them having a characteristic flickering frequency. When the subject focuses on a specific stimulus, its frequency appears in the EEG recordings, allowing to infer which stimulus the subject is focusing on (Bin et al., 2009). Figure 1.21 shows such a paradigm where three stimuli are simultaneously shown on a screen, with 3 different frequencies. Analyzing peaks of the EEG spectrogram allows to detect which part of the screen is being focused on. SSVEP spellers have been successfully designed by using virtual keyboards where each character flickers to a unique frequency (Hwang et al., 2012; Chen et al., 2014). A SSVEP speller based on SSVEP could also work with patients that lost gaze control (Allison et al., 2008).

Since SSVEP are an inherent response of the brain, SSVEP-based BCIs are easy to use and require little training. Moreover their **asynchronous** design does not constrain the user to wait for a cue, who has complete control over when and where to focus. SSVEP-based BCIs also show good information transfer rate and high signal to noise ratio. Some of the limitations include the frequencies of the different stimuli which are constrained by the detection algorithm and screens refresh rates. However alternative paradigms are more suited to applications with multiple degrees of freedom, while requiring training and a more complex synchronized setup (Bin et al., 2009).

1.4.2.3. Slow Cortical Potentials

When recording activity related to movement, movement imagery and other cognitive tasks from the sensorimotor cortex, specific events induce reproducible slow changes in the recorded voltages. Those **Slow cortical potentials** (**SCP**) are voltage shifts time-locked and phase locked to specific events. They occur on longer time scales compared to SSVEP and P300 and can precede or follow an action. For example, a negative SCP called **Bereitschaftspotential** usually precedes self-initiated movements by 500 to 1000ms (Kornhuber and Deecke, 1965). Another SCP called **contingent negative variation** (**CNV**) has been identified when a warning stimulus prepares for another stimulus soliciting an action (Walter et al., 1964). That action can either be motor or cognitive, in particular SCPs have been related to both actual movements and motor imagery (Beisteiner et al., 1995; Cunnington et al., 1996). It seems that SCPs are linked with preparation for action and can characterize specific tasks (Birbaumer et al., 1990).

SCP properties make them good features for BCI systems. With some training a user can learn to perform selection task by performing cognitive actions such as mental arithmetics or motor imagery. However such BCI systems are slow to detect an action due to the low signal to noise ratio and the inherent timing of SCPs. Another drawback of SCPs is the mental load: the cognitive tasks used to control those BCIs need for focus which can be tiring for users and hinders simultaneous control of more than one dimension. Nevertheless, SCPs provide specific information about brain processes and can be combined with other features of brain activity to enhance BCI use (Mensh et al., 2004; Pfurtscheller et al., 2010). A SCP speller can be designed by splitting the alphabet in two groups and presenting both groups successively for a user to select them or not. Then the selected group is split in two halves that are again successively presented for selection, repeating the procedure until one letter remains (Birbaumer et al., 1999).

1.4.2.4. Sensorimotor Rhythms

After inventing EEG in 1924, Berger (1929) described waves in brain recordings that appear when the subject closes its eyes. He coined the term **alpha waves** to describe those 8-12Hz components over occipital electrodes of the recordings and **beta waves** the faster 18-30Hz components with lower amplitude appearing when alpha waves are suppressed(La Vaque, 1999; Kirschfeld, 2005). Over the decades, both alpha and beta waves (also called rhythms) and their link to mental processes were extensively researched along other frequency components of brain recordings. Commonly denominated brain rhythms also include delta rhythms (1-4Hz), theta rhythms (4-8Hz), low gamma rhythms (30-70Hz) and high gamma rhythms (70-200Hz). EEG, MEG and ECoG recordings all display brain rhythms although EEG signals are usually limited to lower frequency rhythms up to 40 Hz.

While the original description of alpha rhythms was done in the occipital lobes, other alpha-like rhythms have also been described in other areas of the brain. In particular, μ-rhythms have been described as 8-12Hz waves in the sensorimotor cortex that are inhibited by somatosensory stimulation and motor behaviors (Pineda, 2005). Beta and μ rhythms recorded in the sensorimotor cortices are called **sensorimotor rhythms** (**SMRs**). Sensorimotor events will inhibit (**desynchronization**) or reinforce (**synchronization**) some SMRs (Pfurtscheller and Aranibar, 1977; Pfurtscheller, 1992). Those **event-related desynchronizations** and synchronizations of SMRs differ depending on their frequency and the triggering event. Some are generalized and triggered by most motor behaviors, while other are specifically lateralized and localized with temporal evolution characteristic to the event (Pfurtscheller et al.,

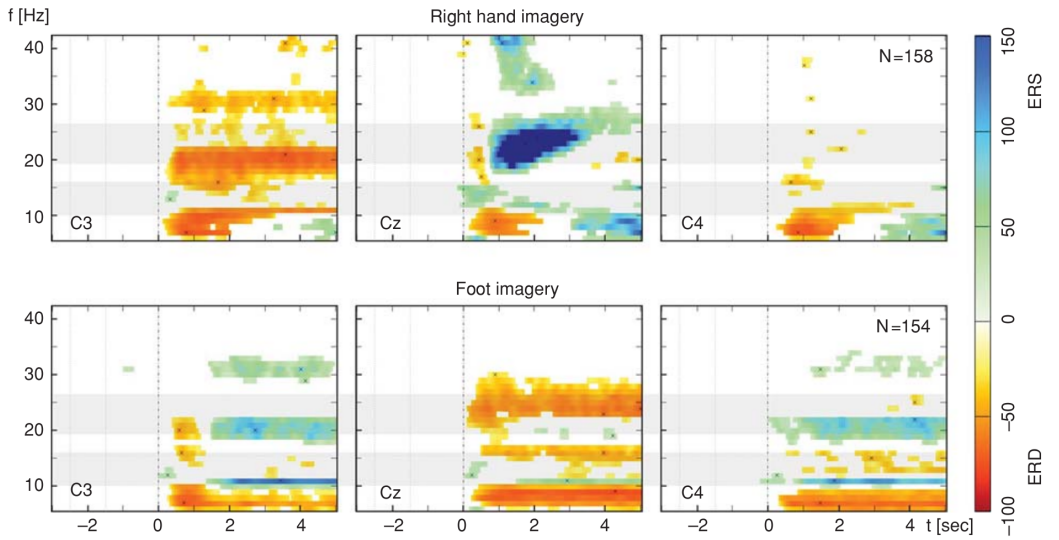


Fig. 1.22.: Antagonist event-related synchronization and desynchronization between foot and right hand imagery. The cue was presented at t=0. C3 shows activity in the left hemisphere and C4 in the right hemisphere, Cz is central. The brain activity is characteristically different betweeen both motor imagery tasks, allowing for efficient classification.
from *Pfurtscheller et al., 2006*

2000). Those properties can be used to classify mental tasks (Fig. 1.22) and use them as controls for a BCI (Krauledat et al., 2008). However regression models are also commonly used in SMR BCI: after calibration with a motor imagery task, the subjects can learn to modulate the amplitude of their SMRs in conjunction with an adaptive decoding algorithm. The resulting control of the BCI strides away from a discrete motor imagery task into a more natural continuous control, like a finely tuned muscle control (Kübler et al., 2005; McFarland et al., 2010).

For some patients, controlling a cursor on a screen would allow to type characters. Such BCIs have been developed to offer a 2D continuous control of a cursor using sensorimotor rhythms (Wolpaw et al., 1991; Kübler et al., 2005). While controlling a cursor can be used for a wide variety of tasks other than communication, its performance for spelling is similar to dedicated P300 spellers.

1.4.3 Invasive BCIs

The non invasive BCIs presented in section 1.4.2 showed very promising results for patients. However none of these BCIs provide a fast and precise enough control that would allow to carry on a normal conversation or perform fluid motor tasks. Moreover, EEG-based communication BCIs require a high level of concentration

(Käthner et al., 2014; Baykara et al., 2016). Invasive BCIs based on intracortical signals seem to allow for better controls and lower cognitive cost as the user gets used to the device after a period of training (Hochberg et al., 2006; Hochberg et al., 2012; Jarosiewicz et al., 2015).

1.4.3.1. Features

Invasive recording techniques can record signals from the activity of a single brain cell (**Single-Unit Activity**, **SUA**), multiple brain cells (**Multi-Unit Activity**, **MUA**) and **local field potentials** (**LFPs**) depending on the technique. SUA, MUA and LFPS have all been used to control BCIs.

Early on, **action potentials** from a single motor cortex neuron have proved to allow real-time control of a one dimensional cursor in monkeys (Fetz, 1969; Fetz and Finocchio, 1971). Further work showed that SUA can encode for a wide range of activity, and MUA can reliably code for tasks with multiple degrees of freedom like 3-dimensional hand trajectory (Georgopoulos et al., 1988). Spikes are short discrete events of about 1ms with high-frequency components. In SUA, modulation of activity is typically measured by the underlying *firing rate* of the neuron's spike trains, which is a measure of the number of spikes in a given time window. When recording multiple neurons at the same time, it is possible to reconstruct individual's neuron's spiking activity using *spike sorting* algorithm that exploit the signature spike shape and amplitude of individual neurons. However this problem is difficult to solve and MUA has been shown to be an effective control signal for BCIs without spike sorting: the signal is bandpassed between 250Hz and 8kHz and spikes are detected with a threshold (Fraser et al., 2009).

Alternatively, **local field potentials** provide information over a localized population of neurons. They can be used to track local sensorimotor rhythms or event-related potentials to control BCIs. But they can also be used as raw features for a decoding system. The raw signals can be low-pass filtered and downsampled, or their spectrograms can be computed on multiple frequency bands.

Both LFPs and spikes are very localized, meaning they provide information about specific brain processes. Spikes can be seen as the output activity of neurons, which provides almost direct insight on brain processes. Carefully placing the electrodes allows to record neural activity directly related to the BCI tasks. For example, recording neurons in the hand motor area would provide natural and efficient control of a robotic hand. Achieving better control parameters also implies the possibility to control more degrees of freedom, allowing for more complex tasks.

1.4.3.2. Decoding Methods

Raw LFPs and spectral features of LFPs are continuous signals by nature. Although the spiking activity recorded from single-units or multi-units is a discrete phenomenon, firing rate on short time windows are also a continuous measure of neural activity. Similar decoding methods can then be used to produce a command signal from neural activity. The decoded BCI command should however be adapted to the recording method. It has been argued that in the motor cortex, SUA better encodes directional tuning to reaching kinematics, while high gamma LFPs and MUA better encode speed, and beta LFPs are suppressed when a movement starts (Perel et al., 2015). Additionally, there's a *time delay* between recorded brain processes and actual physical action, which should be taken into account when building and calibrating a decoder.

BCI typically use either classification methods that produce a discrete output, or regression methods that produce a continuous output. **Discrete decoding** relies on methods such as linear discriminant analysis, support vector machines or neural networks (Moses et al., 2021). **Continuous decoding** relies on methods such as linear regressions (Chapin et al., 1999; Wessberg et al., 2000; Serruya et al., 2002), sometimes with ridge regularization (Collinger et al., 2013; Wodlinger et al., 2015), Bayesian models (Fraser et al., 2009) such as Kalman filters (Guenther et al., 2009; Jarosiewicz et al., 2015), partial least squares regression (Chao et al., 2010; Eliseyev et al., 2012) or neural networks (Chapin et al., 1999; Pandarinath et al., 2018).

1.4.3.3. Invasive Motor BCIs

BCIs using action potentials recorded in the motor cortex were shown to allow effective control of robotic limbs. First by training rats to control a 1-dimensional lever from multi-sites multi-neurons recordings (Chapin et al., 1999), and later on by training monkeys to control robotic arms with 1 to 3 degrees of freedom (Wessberg et al., 2000).

Such invasive BCI using microelectrode array were shown to allow people with long standing tetraplegia to control a robotic arm and hand to perform 3-dimensional reach and grasp movements, although with a limited accuracy (Hochberg et al., 2012). A linear discriminant analysis and a Kalman filter were trained to respectively predict the hand state and velocity from MUA. One subject was famously able to use it to drink from a cup with a straw, although with some struggle. Further work increased the number of degrees of control of robotic arms from 4D to 7D by adding

separate control for rotations (3D), translations (3D) (Collinger et al., 2013) and to 10D by decomposing the grasp control into 4 independent dimensions (Wodlinger et al., 2015).

Combined with 36 implanted percutaneous electrodes stimulating a participant's arm, a similar invasive BCI using 2 microelectrode arrays in the hand motor cortex allowed for the patient to move his own arm (see Fig. 1.23) and perform basic tasks such as drinking from a mug (Ajiboye et al., 2017). That opens the possibility for paralyzed patient to recover some control over their own limbs. Another study focuses on developing a BCI controlling an exoskeleton using ECoG grids implanted over the upper limb sensorimotor areas of a participant with tetraplegia (Benabid et al., 2019). The BCI provides a control over the arms of the exoskeleton as well as a switch that controls walking.

Invasive motor BCIs rely on visual feedback for the subject to learn how to control it, while control of actual limbs rely primarily on somatosensory feedback. Sensory neuroprosthesis have been found to restore tactile sensation by performing microstimulation of the somatosensory cortex, in particular feeling of pressure and contact location (Flesher et al., 2016). A bidirectional BCI using both a decoding module and a somatosensory feedback module was recently tested for the first time in humans (Flesher et al., 2021). The patient was able to control robotic arm and hand with 5 degrees of freedom using 2 microelectrode arrays implanted in his left motor cortex. Two additional microelectrode arrays implanted in the cutaneous somatosensory cortex were used to simulate palm and fingers somatosensory feedback from grasping with the robotic hand. The evoked feedback was found to improve speed at picking up objects with the robotic arm.

1.4.3.4. Invasive communication BCIs

In parallel with the development of BCIs for controlling of robotic arms, the same methods have been applied to the control of virtual cursors. Monkeys implanted with utah arrays in the primary motor cortex learnt to freely control a 2D cursor on the computer screen. During a training phase the cursor was controlled by hand using a lever, before controlling it exclusively using the BCI. A reaching task where the cursor had to be moved on targets was performed almost as fas as using hand control by one of the monkeys (Serruya et al., 2002). A participant implanted with a 96 microelectrode array in the primary motor cortex trained to control a 2D cursor with limited accuracy with simple applications such as controlling a television (Hochberg et al., 2006).

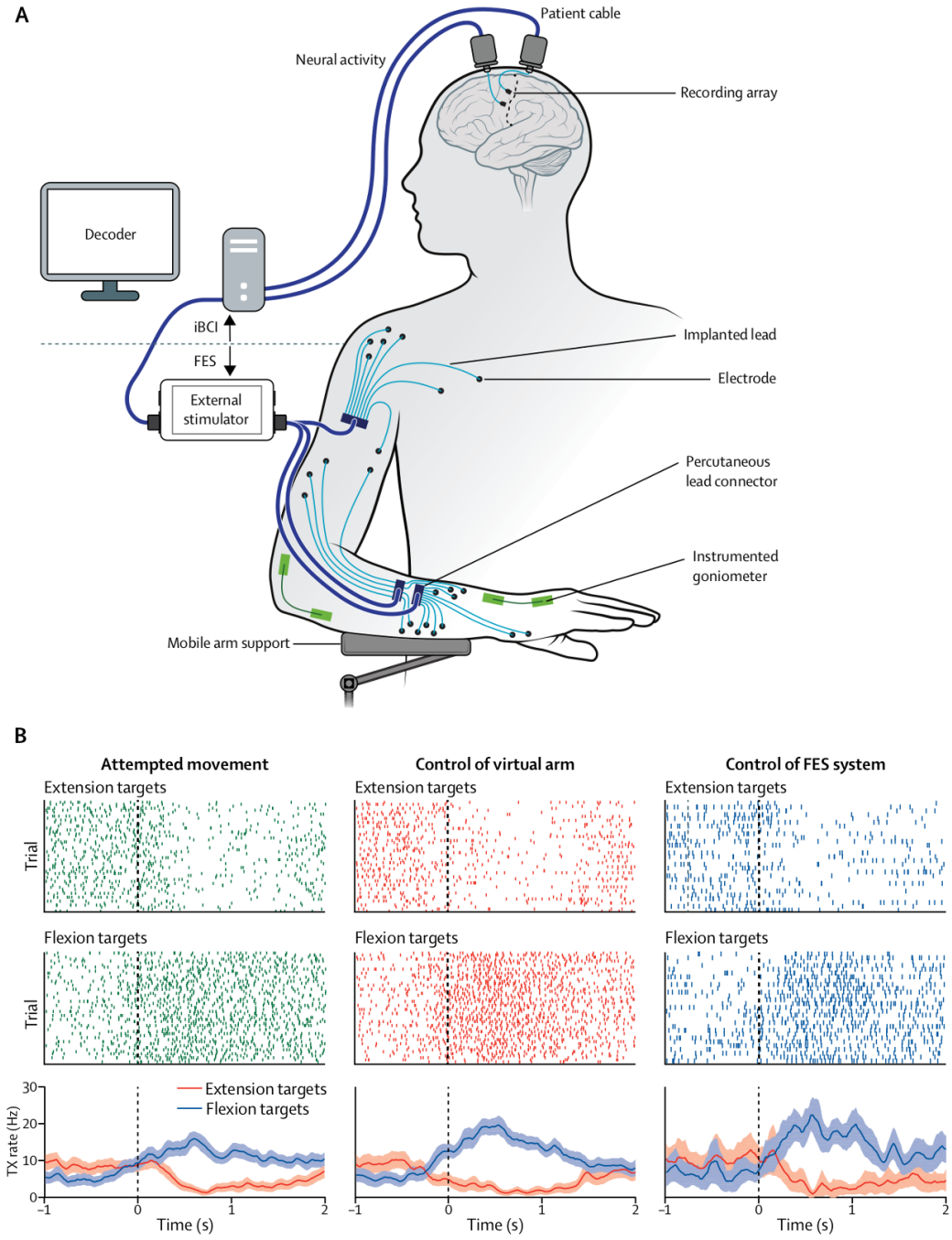


Fig. 1.23.: Invasive BCI combined with functional electrical stimulation to control the participant's own arm. A. Neural activity recorded from two microelectrode arrays implanted in the motor cortex was used to control the amount of functional electrical stimulation (FES) and the elevation of the mobile arm support B. Example raster plots showing the timing of threshold crossings (top rows) and the average threshold crossing rates (bottom row) of a single channel tuned to wrist flexion and extension during a single-joint wrist movement task. The dotted line at t=0 indicates the presentation of the target movement. This channel records more threshold crossings when flexion targets are presented and has similar tuning properties during all three experimental conditions: attempted movement of the paralyzed arm (left column), control of a virtual arm in a virtual reality game (middle column) and control of the FES system (right column).
adapted from *Ajiboye et al., 2017*

BCIs for controlling a computer cursor were further studied by providing both a free control over 2 dimensions and a click control. Such a cursor allowed 4 participants with tetraplegia implanted with microelectrode arrays in the hand and arm motor cortex to type on a virtual keyboard (Jarosiewicz et al., 2015). A Kalman filter was used for the cursor control, while a linear discriminant analysis was used for detecting the click control. An adaptive recalibration was set up to correct biases and adjust decoding during use without requiring pauses for calibration. Further work adding an HMM (hidden Markov model) based click decoder reached a performance for virtual typing up to 40 characters per minutes (Pandarinath et al., 2017). Lately, deep learning - more specifically variational autoencoders adapted to sequences using recurrent neural networks - have been argued to improve the cursor control by extracting robust representations of neural activity across trials on longer time spans (Pandarinath et al., 2018).

Very recently, a communication BCI was built on decoding handwriting neural activity and converting it to characters on a computer screen (see Fig. 1.24). A participant with tetraplegia was implanted with two microelectrode arrays in the hand motor cortex. Multi-unit threshold crossing rates of individual characters activity were classified in real-time using a recurrent neural network classifier at up to a 94.6% raw accuracy during online use and 99.1% offline accuracy using a Viterbi-based autocorrect. This method achieved state of the art performance with up to 90 characters per minute which is close to regular typing speed on a smartphone of about 115 character per minute (Willett et al., 2021).

1.4.3.5. Speech BCIs

The communication BCI systems we presented so far all use an indirect design, they involve mechanisms not related to speech to somehow perform a communication task. Drawbacks of this strategy are: 1. it is not intuitive 2. it requires fully motor resources unrelated to speech while using the BCI, thus limiting the ability of the subject from simultaneously communicating and performing another motor task. An alternative would be to decode speech related neural activity directly from speech brain areas using invasive electrodes. Real-time vowel production was achieved by continuously controlling a formant synthesizer using a neurotrophic electrode implanted in the left ventral premotor cortex (Guenther et al., 2009). A Kalman filter was used to predict the two first formant's velocities and positions from spiking activity (see Fig. 1.25). The synthesizer provided feedback under 50ms with an accuracy of 70% after training. Recently, a BCI using high density ECoG allowed to predict words out of a 50-words dictionary by classifying neural activity in near

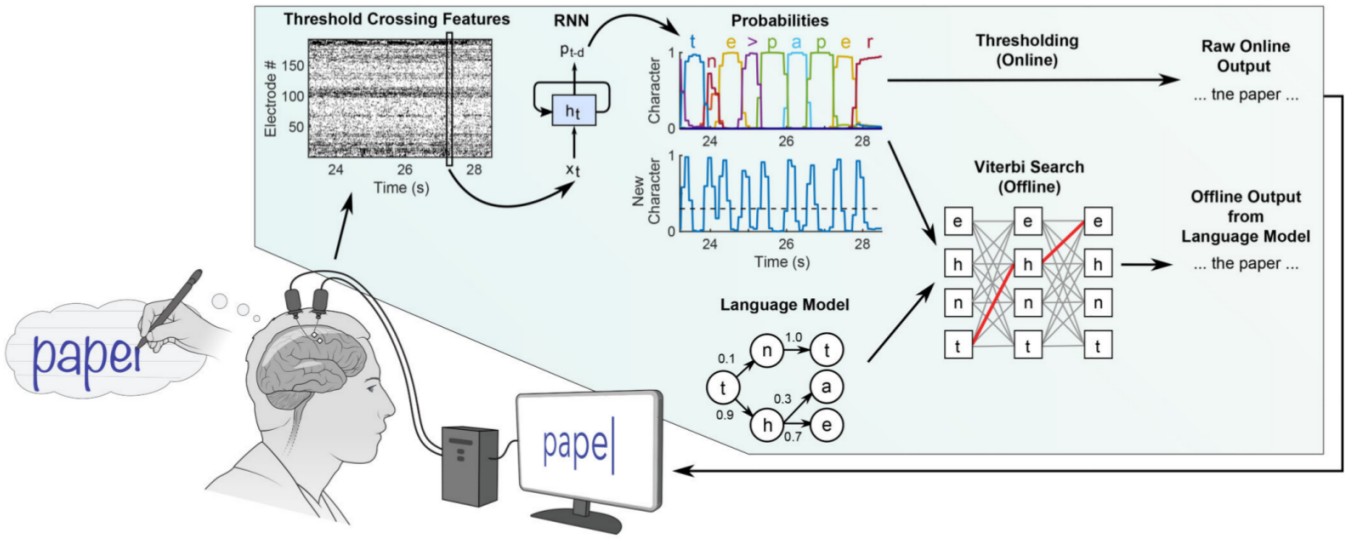


Fig. 1.24.: Neural decoding of attempted handwriting in real-time. A recurrent neural network (RNN) was trained to predict the probability of a character from MUA of handwriting with a 1s delay. For real-time use, the character were decoded when their probability crossed a threshold while a language model was used to automatically correct predictions on an offline setting. From *Willett et al., 2021*

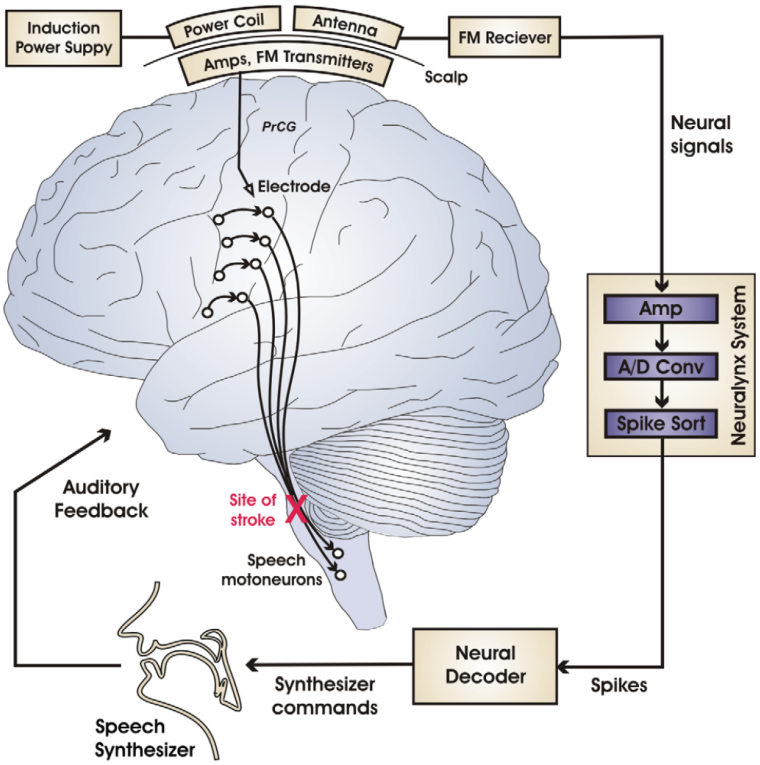


Fig. 1.25.: Schematics of BCI for real-time formant synthesis. Speech motor neurons are represented by black circles and axonal projections. Speech formants were decoded from recorded signals to synthesize vowels and provide real-time audio feedback. From *Guenther et al., 2009*

real-time (Moses et al., 2021). A language model and a decoding neural network were combined to achieve a speed of 15 words per minute, with 25% word error rate. Finally, a speech BCI allowing real-time closed-loop synthesis of continuous speech was tested with a participant implanted with stereotactic electrodes (Angrick et al., 2021). Log-mel spectrograms were decoded from neural activity using linear regression. After training on actual speech, the participant controlled the BCI with immediate feedback using whispered and covert speech, although without reaching intelligibility.

Invasive BCIs show promising results to improve speed, ease of use and accuracy of communication BCIs. In order to raise performance to a natural level, a continuous control of a speech synthesizer seems appropriate. It would require a system able to continuously control multiple degrees of freedom in real-time, which would not be possible with a non invasive method. These elements motivate the further investigation of invasive speech BCIs for natural speech.

1.5 Speech decoding

Although promising, invasive speech BCIs do not allow for restoration of natural speech yet. Contrary to their non invasive counterparts, invasive speech BCIs require surgical acts and clinical monitoring. The benefit they can offer to the patients so far does not justify to take medical risks. Yet it would be impossible to develop such BCIs without access to invasive brain recording. In order to prepare the way for **online** trials where the patients could learn to control the BCI, research have been focusing on decoding speech from **offline** invasive recordings.

1.5.1 Data

In order to study decoding of speech for speech BCIs, one needs to record neural activity during speech production. In this section we describe in which conditions invasive recordings of speech can be done and what are the possible speech tasks.

1.5.1.1. Patients

Some rare patients are implanted with invasive BCI for research applications, however the vast majority of speech invasive recordings are performed in parallel with

a necessary medical act: 1. during awake brain surgery, or 2. during intracranial monitoring of epilepsy.

In order to treat some neurological disorders such as brain tumors, some patients need to undergo **awake brain surgery**. After opening of the skull, the patients are awakened and asked to performed specific tasks related to the brain areas surrounding the tumor. By applying small electrical stimulations to those brain areas, the surgeon maps how that interferes with the cognitive tasks, effectively establishing a mapping of the patient's brain around the tumor. This mapping allows the surgeon to remove the tumor while preserving the patients cognitive abilities. This procedure can be done without hurting the patient, as there is no pain receptors in the brain. During the surgery, electrodes can temporarily be positioned on or implanted in the patient's brain during a speech task. Data is recorded only once with these patients, in the context of a surgical room. Due to anesthesia and surgery, the patients can have issues to perform the tasks and can present some speech impairments.

Some patients suffering from epilepsy resisting to pharmacological treatment require surgery. In case of epilepsy originating from a focal source, invasive methods can be required to localize the source of the epilepsy prior the surgery. Implanted patients are typically monitored in the hospital for a few days up to two weeks (Haut et al., 2002) in order to record seizures. During **intracranial epilepsy monitoring**, speech neural activity can be recorded from patients with relevant electrode placement. Over the duration of the monitoring, recording sessions can be spread over multiple days. That gives an opportunity to let the patient use a BCI and train over separate sessions. Recording sessions can also be longer than during awake surgery, and in better conditions for the patients.

1.5.1.2. Speaking condition

Neural activity reflects cognitive tasks. In particular, different speech tasks will have different neural substrates as described in section section 1.3.3.1. When developing speech decoding methods for ALS or locked-in patients, it is important to consider the effect of physical inability to speak on neural activity. Indeed, methods are often first developed on healthy patients recordings. The recorded task should be carefully designed so that decoding methods may later translate to patients. Different speech production conditions have to be taken into account: 1. overt speech production, 2. silent articulation, 3. inner/covert speech.

Overt speech is the most natural condition of speech. Patients with minimal or no control of articulators could still attempt overt speech although they would not have sensory feedback from muscles nor audio feedback. Healthy subjects and paralyzed patients should share similar cortical activity, especially considering that internal sensorimotor models would still work (see section 1.3.3.3).

It has been debated however that **covert speech** might be a better condition to decode speech in normal subjects for further application in paralyzed patients (Perrone-Bertolotti et al., 2012). Covert speech might be defined as imagined speech or sometimes imagined articulation, although those conditions differ. Although it remains not well understood (Martin et al., 2018), covert speech shares some neural substrate with overt speech (Martin et al., 2014).

The relevant recording sites for covert and overt speech cortical activity are very similar (Bocquelet et al., 2016a). Hypothetically a decoding framework trained on overt or silent speech neural activity might also translate to decoding covert speech activity. In practice however, decoding of overt speech typically yields better performance than covert speech (Martin et al., 2014; Martin et al., 2016). It has been argued that covert speech signals recorded with ECoG might provide less information about motor control of articulators (Pei et al., 2011).

In the context of a closed loop speech BCI control, healthy patients could not resort to attempted overt speech like locked-in patients because they need to prevent vocalization by engaging inhibitory mechanisms that are not engaged by paralyzed patients attempting to speak. They would therefore have to control the BCI using covert speech. It is indeed possible that overt speech decoders not only decode neural activity related to speech production but also to speech perception, as the subjects also hear the feedback from their own voice. This is especially true with decoding models that use a time buffer to capture "future" neural activity. Comparatively, a patient using a natural speech BCI would only have audio feedback from the speech synthesizer, which implies that decoding models would not benefit from decoding perceived speech. In order to control for this bias, Anumanchipalli et al., 2019 investigated decoding from a **silent articulation** task. They showed that speech can still be decoded from silent articulation neural activity, although not nearly as well as overt speech. Silent articulation should be closer to attempted overt speech than covert speech as it ensures actual control of speech articulators with somatosensory feedback, but no actual sound production thus laryngeal activity and feedback.

1.5.2 Decoding framework

Speech representations used for speech synthesis include on one side discrete linguistic representations such as phonemes, words or syllables, and on the other side continuous abstract representations like mel cepstrum (see section 1.2.6). Similarly a decoder of speech from neural activity should aim at decoding either discrete features of speech, or continuous features of speech. Speech could then be synthesized from the decoded features by an appropriate synthesis method (Fig. 1.26).

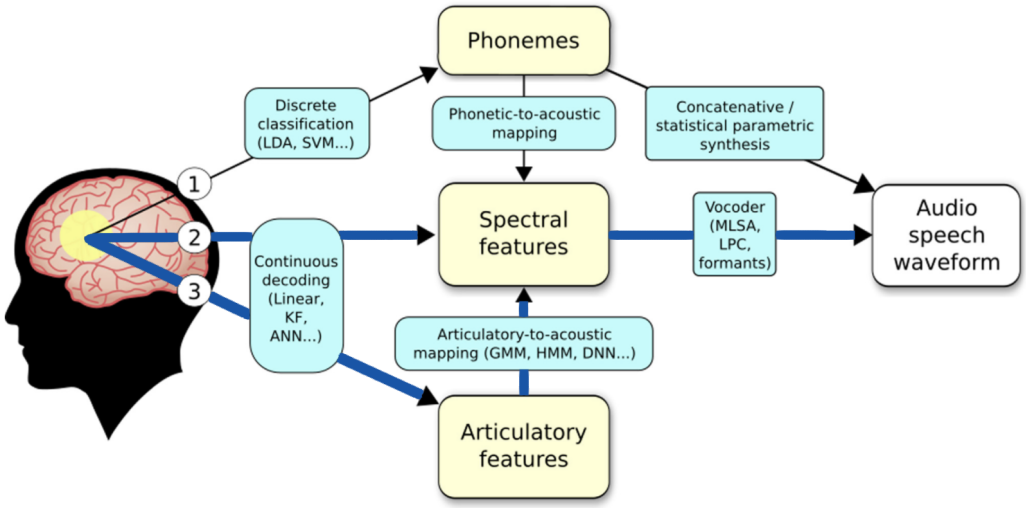


Fig. 1.26.: Continuous and discrete speech decoding frameworks. The approaches compatible with natural continuous speech BCI are marked by thick blue lines. adapted from *(Bocquelet, 2017)*

1.5.2.1. Discrete decoding

When decoding discrete features of speech, the decoder has to perform a classification task to associate a meaningful speech representation to a chunk of neural activity. Words, sentences, phonemes and syllables have all been investigated features for discrete decoding of speech.

Words have been successfully decoded with good accuracy from ECoG recordings in an epileptic patient (Kellis et al., 2010). This approach requires to build a small dictionary of words to classify. The bigger the dictionary, the more difficult it is to decode words correctly. Limiting possible vocabulary to 10 to 100 words would impose a strong restriction on the communication.
Such a vocabulary size would be however enough to represent all the phonemes of

a given language. By decoding a similar number of **phonemes** instead of words, it would be possible to build any sentence in a given language. Decoding of 35 American English overtly produced phonemes from ECoG recordings of 4 epileptic patients reached an overall performance of about 20% accuracy, with the best subject achieving 36% accuracy (Mugler et al., 2014). Vowels have been found to be encoded both in individual neurons and at neural population levels in 11 epileptic patients implanted with intracranial depth electrodes terminated with microwires. Five vowels were classified with 93% accuracy from intracranial depth electrodes terminated with microwires recordings of 11 epileptic patients (Tankus et al., 2012), and 59% from intracranial recordings of 6 epileptic patients using a combination of microneedles and macroelectrodes (Ibayashi et al., 2018).
Both decoding of words and phonemes can be combined with a **language model** to improve decoding. Decoding words from ECoG recordings of 7 patients using a statistical language prior yielded down to 25% error rate for a 10-word dictionary and 50% error rate for a 21 phone dictionary (Herff et al., 2015). A recent study also showed a clear improvement from using a language model, achieving 25% error rate on a 50 words dictionary in an online setting (Moses et al., 2021).
Some work investigated the decoding of entire **sentences** from speech production ECoG recordings (Moses et al., 2019) or from speech perception (Moses et al., 2018).

1.5.2.2. Continuous decoding

Speech can be described by the continuous trajectories of speech articulators or spectral features of speech sound. Since brain cognitive processes of speech production include both motor commands to control articulators and an internal auditory model of speech, both articulatory and acoustic representations of speech are good candidates for speech decoding from neural activity. Decoding continuous features of speech requires to perform regression to map neural features to speech features.

Multiple continuous **acoustic features** of speech have been subject to decoding from neural activity. Spectrograms of perceived words and sentences from a limited set have been reconstructed from the auditory cortex by linear regression (Pasley et al., 2012). Spectrograms of produced speech have also been decoding in real-time from ECoG recordings of the temporal areas, with a real-time synthesis of the speech waveform using Griffin-Lim algorithm (Herff et al., 2016). Other studies focused on decoding vocoder features of speech including voicing, f0, aperiodicity and spectral envelope, showing a better performance compared to decoding spectrograms of speech (Akbari et al., 2019). Finally formants trajectories were decoded to control a

formant synthesizer in real-time by a locked-in patient, effectively producing vowels with good accuracy (Guenther et al., 2009). Although the modalities are different, an EEG study of a real-time formant BCI showed that adding a visual feedback of the decoded formant frequencies could help control the formant synthesizer (Brumberg et al., 2018).

The number of acoustic features required to synthesize arbitrary complex speech is around 25 to 30 for a vocoder-based synthesis. Spectral features of speech are a good condensed representation of acoustic speech as shown in section 1.2.5, however articulatory trajectories also encode speech well with a potential compression down to 7 to 14 features. Decoding an intermediate **articulatory representation** chained with and articulatory-to-acoustic mapping could allow to control a speech synthesizer with few parameters that would perform better than a formant synthesizer to produce consonants (Bocquelet et al., 2016a). In recent work achieving high quality reconstruction of speech from ECoG recordings, decoding an intermediate articulatory representation later mapped to vocoder acoustic features improved performance over direct decoding of vocoder features (Anumanchipalli et al., 2019). Some studies have been tracking articulatory trajectories in sensorimotor cortex. Since synchronized recordings of ECoG and electromagnetic articulography is not possible, articulatory trajectories have to be estimated from speech (Conant et al., 2018; Chartier et al., 2018).

1.5.3 Decoding methods

1.5.3.1. Neural features

Despite the success of Neurotrophic electrodes designed for long term implantation and recording of action potentials (Bartels et al., 2008; Guenther et al., 2009), many studies on speech decoding from invasive recordings focus on Electrocorticography (ECoG). Compared to electrodes recording single or multi units activity, ECoG records local field potentials over a wider range of the cortex. It has been widely used for monitoring of epileptic patients and is compatible with brain surgery for recording of speech datasets. Aside from practicalities, ECoG has been shown to record speech and motor activity (Leuthardt et al., 2004). Alternatively, stereotactic-EEG which is also used for epilepsy monitoring has also been considered for speech BCIs in a few studies (Herff et al., 2020; Meng et al., 2021). Lastly, it was showed that combining multiple intracranial recording signals such as ECoG local field potentials and single unit activity can improve speech decoding from cortical activity (Ibayashi et al., 2018).

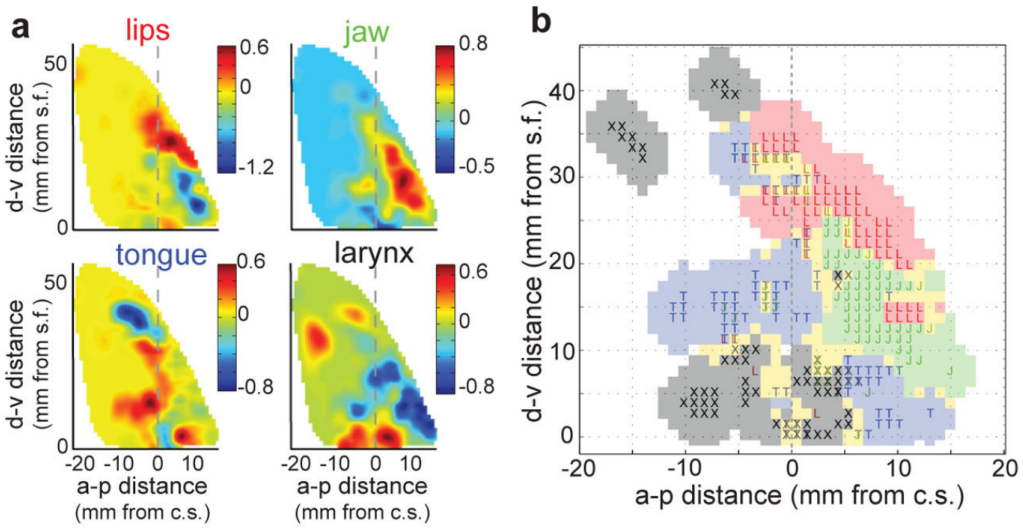


Fig. 1.27.: Spatial Representation of Articulators. **a**, Localization of lips, jaw, tongue, and larynx representations in the ventral sensorimotor cortex (vSMC). The anterior-posterior (horizontal) axis is measured from the central sulcus and the dorsal-ventral axis from the Sylvian fissure (vertical). **b**, Functional somatotopic organization of speech articulator representations in vSMC. Lips (L, red); jaw (J, green); tongue (T, blue); larynx (X, black), mixed (Gold). Letters correspond to locations based upon direct measurement, shaded rectangles correspond to regions classified by k-nearest neighbor.
from *Bouchard et al., 2013*

ECoG grids are typically placed over the sensorimotor, premotor and auditory cortex as well as Wernicke's area for decoding speech. While all these areas contribute to decoding produced speech (Lotte et al., 2015), perceived speech has been decoded from auditory cortex (Pasley et al., 2012) and articulatory trajectories have been shown to be encoded in the sensorimotor cortex (see Fig. 1.27) (Bouchard et al., 2013; Conant et al., 2018) .

Aside from spectral features corresponding to event related potentials in alpha, beta and gamma bands, higher gamma frequencies have been shown to carry information for speech decoding from 70Hz up to 200Hz (Leuthardt et al., 2004; Bouchard et al., 2013; Moses et al., 2018; Anumanchipalli et al., 2019) or even up to 300Hz (Crone et al., 2001).

1.5.3.2. Machine learning methods

Traditional decoding schemes focus on **linear methods**. Discrete decoding of speech has been done by linear classifiers (Mugler et al., 2014) and support vector machines (Martin et al., 2016). For continuous decoding of speech, linear regressions (Herff et al., 2016) and kalman filters (Guenther et al., 2009) have been used. As linear models are fast to train and very fast to compute, they are suitable for real-time use. Moreover, BCIs based on linear models are believed to offer controls to the users that are easy to learn.

Some recent works published during the time frame of my PhD now include **deep learning** methods with promising results. Already several EEG-based BCIs have been using recent machine learning techniques (Schirrmeister et al., 2017; Lotte et al., 2007), which could also be used also for intracortical recordings. Recent work applies deep learning methods to classify auditory sentences in real time, beating other decoding methods (Moses et al., 2018). Deep neural network also establishes state of the art performance in syllables decoding from sensorimotor cortex recordings, showing data is key to increase the performance (Livezey et al., 2018). Recent approaches also used deep learning to decode continuous features of speech, beating linear methods (Angrick et al., 2018; Anumanchipalli et al., 2019).

The interest for a deep learning based speech BCI is corroborated by current results on a deep neural network based motor BCI showing fast response, high accuracy, with a performance sustained on more than a year (Schwemmer et al., 2018). However deep learning methods require large amounts of data to train decoding models, which is not easy to gather for invasive speech recordings. Moreover, the

current neural networks used in state of the art decoding of continuous speech are not compatible with real-time as they process whole sentences at a time. It remains unclear whether such DNN-based decoder processing whole sentences at a time ensure generalization capabilities beyond the set of speech items they have been trained on. This is important for free speech synthesis.

2 General objectives of the thesis

This thesis aims at developing machine learning techniques for the decoding of speech from invasive recordings of neural activity, with a focus on linear methods. It is part of a long-term attempt to build a BCI that would allow its user to produce natural speech in real-time.

The first objective of this thesis is to investigate linear methods for direct decoding of acoustic speech features from neural activity in an offline setting. This requires to develop efficient processing of existing neural and acoustic recordings of speech production. In particular, this work examines compact spectral representations of neural activity using feature selection and features reduction techniques. Chosen speech representations focus on one side on F0 and mel cepstrum for a vocoder-based synthesis, and on the other side on F0 and formants for a formant-based synthesis. Those methods should all be compatible with real-time use in order to be included later in a speech BCI.

The second objective of this thesis is to compare direct decoding of acoustic coefficients from neural activity with an indirect decoding of acoustic coefficients through an articulatory representation. Indirect decoding requires to decode articulatory trajectories from neural activity, and predict acoustic speech from decoded articulatory trajectories. Decoding of articulatory trajectories focuses on the same linear methods used for direct decoding of acoustic speech, while articulatory-to-acoustic synthesis focuses on neural network based approaches.

3 Methods

3.1 Data

This work required synchronized recordings of audio, articulatory trajectories and neural activity of speech. The chosen methodologies were Electromagnetic Articulography (EMA) for recording articulatory trajectories, and Electrocorticography (ECoG) for brain activity. As simultaneous EMA and ECoG recordings are not practically compatible, separate EMA and ECoG datasets were considered.

3.1.1 EMA datasets

3.1.1.1. BY2014

BY2014 (Bocquelet et al., 2016b) is a large articulatory-acoustic corpus containing the recording of vocal tract movements and simultaneous audio signals in one french male speaker reading 676 short sentences including isolated vowels and VCVs (vowel-consonant-vowel sequences like *'apa'*, *'iti'*,...). ElectroMagnetic Articulography (EMA) was recorded with 9 3-Dimensional sensors at 100 Hz positioned on lips corners, upper and lower lips, tongue tip, back and dorsum, soft palate and jaw (actually front teeth). Head movements were removed from the recordings so that articulatory trajectories are describing movements relatively to the head. The corpus therefore consists of 27 articulatory features and the matching audio recording. Recording sites and example trajectories projected on the midsagittal plane can be observed in Fig. 3.1.

3.1.1.2. MOCHA-TIMIT

The MOCHA corpus (Wrench, 1999) contains recordings of one male and one female native english speakers reading a set of 460 short sentences. Only the male speaker was investigated in this work. The sensors recorded trajectories of 7 articulators: jaw, upper and lower lips, soft palate and tongue tip, blade and dorsum. EMA

recordings were sampled at 500 Hz, corrected from head movements and projected in the midsagittal plane, yielding 14 articulatory features.

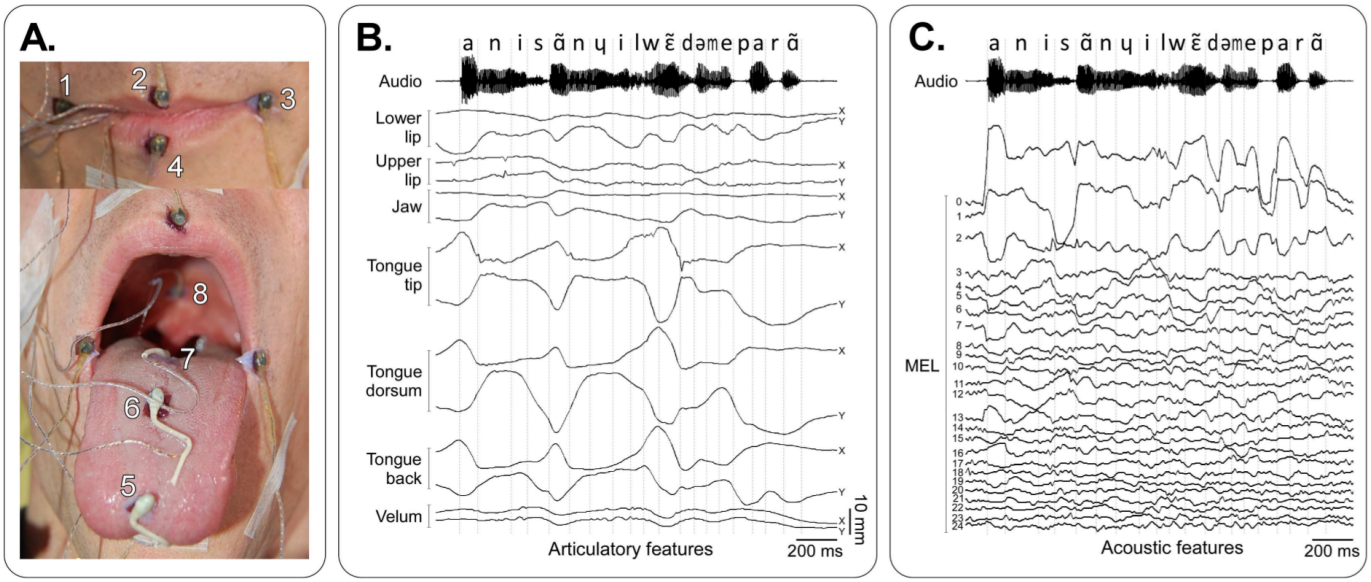


Fig. 3.1.: Articulatory and acoustic data on BY2014
A. Display of the acquisition sensors. Sensors 1,2,3, and 4 record the trajectories of the lips; sensors 5,6, and 7 record the trajectories of the tongue tip, dorsum, and back; sensor 8 records the velum. A 9th sensor was glued at the base of the incisive to account for jaw movements.
B. Articulatory trajectories projected in the midsagittal plane of the sentence "*Annie s'ennuie loin de mes parents*"
C. Acoustic features of the sentence "*Annie s'ennuie loin de mes parents*"
from *Bocquelet et al., 2016b*

3.1.1.3. PB2007

PB2007 contains recordings of one male native french speaker reading 1109 sentences including only 117 short sentences, the rest being VCVs, CVCs, and isolated vowels. The corpus features 6 positions of the vocal tract: lower and upper lips, jaw and tongue tip, back and dorsum. EMA were recorded at 200 Hz, head-corrected and projected in the midsagittal plane, yielding 12 articulatory features.

3.1.2 ECoG datasets

In order to later combine articulatory trajectories and neural activity, two patients implanted with ECoG arrays were recorded reading sentences from the BY2014

dataset. Some additional data of a patient reading sentences from MOCHA-TIMIT dataset was provided to us by Prof. Edward Chang, UCSF.

3.1.2.1. Patients

Both patients were participants of the Brainspeak clinical trial (NCT02783391) approved by the French regulatory agency (DMDPT-TECH/MM/2015-A00108-41) and the local ethical committee (CPP-15-CHUG-12). The first participant (P2) was a 42-year-old patient undergoing awake brain surgery for tumor resection. The second participant (P5) was a 38 year-old female implanted for 7 days as part of a presurgical evaluation of her intractable epilepsy. Both patients gave their informed consent to participate in the study.

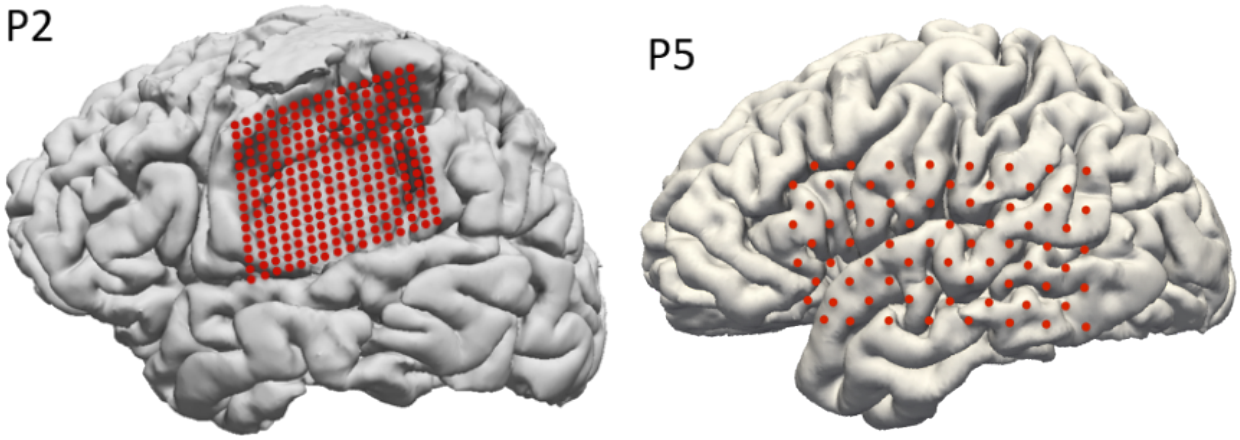


Fig. 3.2.: **Electrode placement for P2 and P5.** A 256-electrode array was positioned over the left sensorimotor cortex of P2 (left) during awake surgery. A 72-electrode array was implanted in P5 (right) largely covering the left hemisphere.

3.1.2.2. Recording

Brain activity from participant **P2** was recorded during awake surgery in the operating room just before tissue 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 3.2, left). Ground and reference electrodes were integrated on the back side of the array. Audio and brain signals were recorded and amplified synchronously at 10 kHz.

Brain activity from participant **P5** was recorded in her room at the hospital. This participant was implanted with a 72-electrode array (PMT Corp., USA) covering a

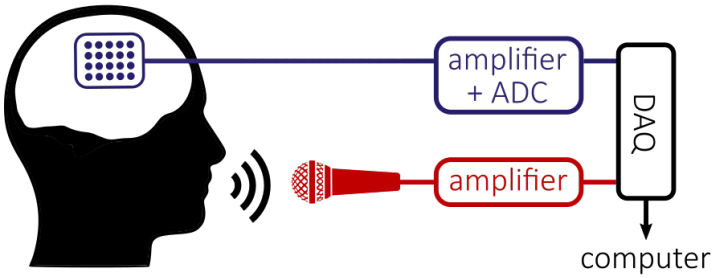


Fig. 3.3.: **Representation of the recording setup for P2 and P5.** The neural data stream is represented in blue and the audio data stream in red. The Analog-to-Digital Conversion (ADC) is performed in the Data Acquisition System (DAQ) for the audio signal and in the front-end amplifiers (FEA) for the neural signal.
from *Roussel et al., 2020*

large portion of her left hemisphere as well as a 4-electrode strip (PMT Corp., USA) over the left ventral temporal lobe (figure 3.2). One electrode of the strip was used as the reference and one as the ground. An additional 96-electrode micro Electrode Array was also implanted in the patient's cortex but was not used in this work. The audio and brain signals were recorded and amplified synchronously at 30 kHz.

3.1.2.3. Tasks

Both participants were asked to read aloud a set of short French sentences from BY2014 dataset (see section 3.1.1.1). Out of the 676 sentences of BY2014, P2 read out loud 118 of them during his awake surgery, including only short natural sentences. As the recording was only possible during surgery, P2 recordings were limited to one session. P5 on the other hand could participate to multiple sessions spread over 4 days of recordings.

During the 4-day experiment, P5 participated to both closed loop and open loop tasks. During **open loop** experiments, P5 read sequences of vowels and short sentences from BY2014 without any audio feedback. Depending on the recording sessions, P5 produced each sentence following multiple speaking conditions: first reading, then repeating the same sentences, and lastly covertly repeating it again before saying "*ok*" when done. Both *read* and *repeat* conditions required to speak out loud, but the written sentence was only displayed on the screen during *read* condition. For the *covert* condition, P5 was asked to imagine repeating the sentence once more,

without actually producing speech or moving the articulators, and without seeing it on screen, .

For this work, only open loop recordings of the three first days of experiments were used. During day 1, P5 read, repeated and covertly repeated 97 sentences including 4 repetitions of 4 vowel sequences (*'a, i, ou'*; *'u, é, è'*; *'e, o, an'*; and *'on, in'*). During day 2, P5 read 141 sentences, including 6 repetitions of 4 vowel sequences. During day 3, P5 read, repeated and covertly repeated 153 sentences, including 7 repetitions of 4 vowels sequences. This amounts to a total of 391 read sentences, 250 repeated sentences and 250 covert sentences.

3.1.2.4. Additional data: EC61

This additional dataset was provided by Prof. Edward Chang, University of California, San Francisco.

The EC61 dataset contains synchronized audio and ECoG recordings of a 30-year-old english-speaking female patient implanted for epilepsy monitoring. She was implanted with a 256 electrodes array broadly covering her right hemisphere (Fig. 3.4), which was found to be her dominant speech hemisphere. She read a total of two sets of 460 sentences from MOCHA-TIMIT, a mix of passages from famous stories, as well as a set of mimed sentences. The data that was provided to us was restricted to 9 recording blocs from MOCHA-TIMIT, amounting to a full set of 460 sentences.

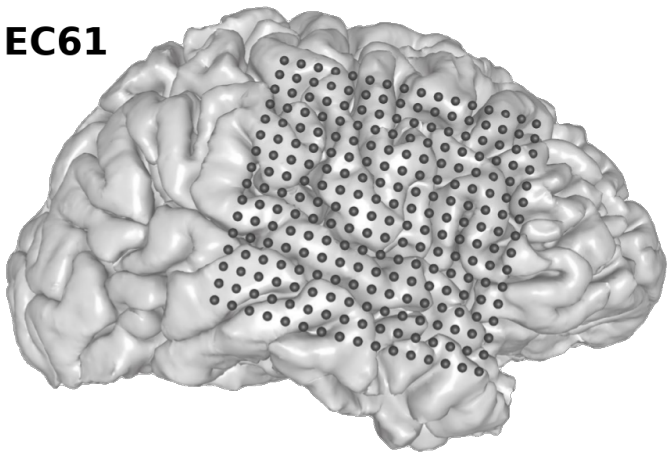


Fig. 3.4.: Electrode placement for EC61. A 256-electrode array was implanted in a right-lateralized patient for epilepsy monitoring, broadly covering its right hemisphere. adapted from *Anumanchipalli et al., 2019*

3.1.2.5. Annotation

The P2 dataset was already processed and annotated by a previous PhD student, however this was not the case for P5 and EC61. A labelling graphical interface was developed in Matlab with Philémon Roussel (PhD student at the time) in order to facilitate the annotation of ECoG datasets (Fig. 3.5). All sentences were manually inspected one by one to annotate the condition, transcription, phonetic transcription and if necessary to discard failed attempts or trials with noisy backgrounds. Sentences were automatically cut using a speech envelope detection so that only 500ms of silence remains before and after speech, although some manual adjusting was necessary. The annotated speech conditions were *read*, *repeat*, *covert*, and *rest* that labelled resting intervals in between trials.

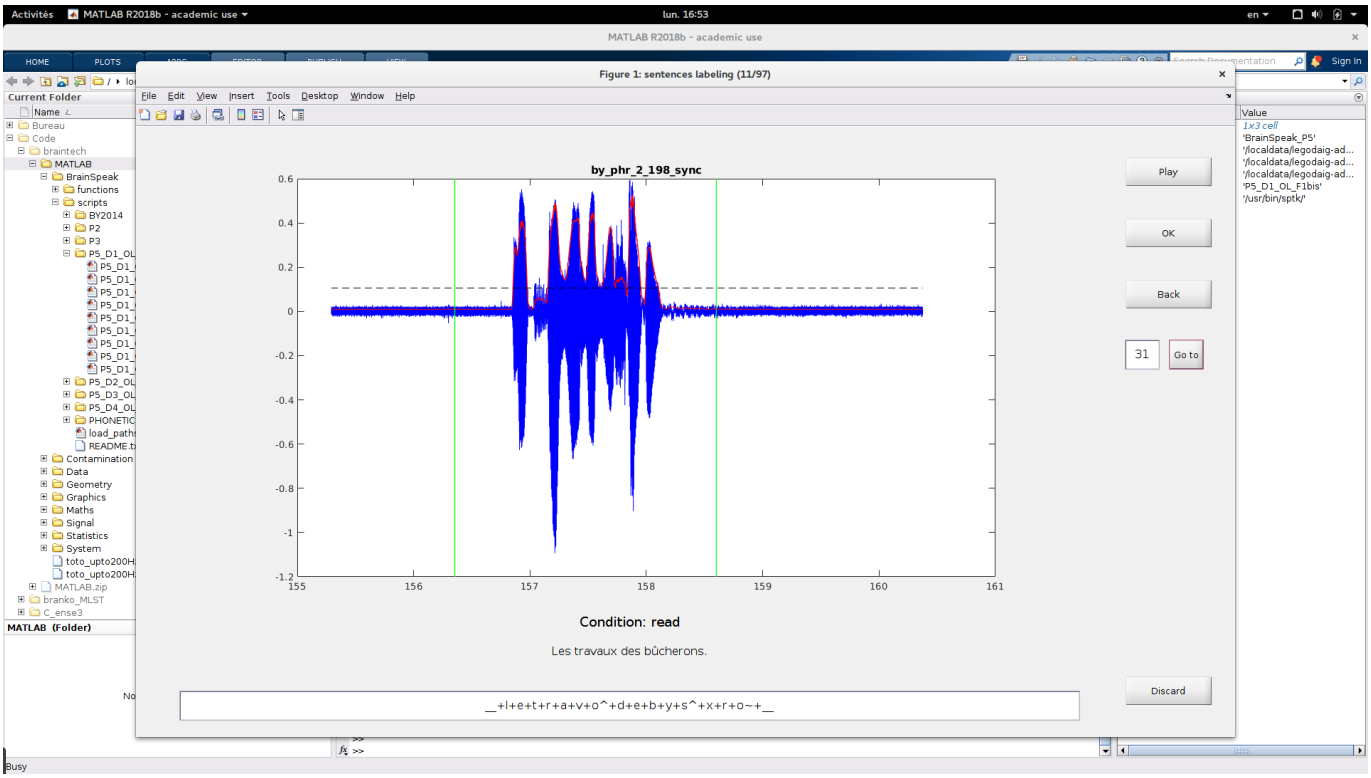


Fig. 3.5.: Labelling interface. Editing of a trial of P5 dataset. The segmentation of the trial (green vertical bars) is automatically set by detecting speech envelope detector (red) above a threshold (dashed line). Segmentation and annotations can be manually edited, and bad trials can be discarded.

3.2 Neural Data Processing

3.2.1 Preprocessing

Artefacts such as line noise were removed from neural signals using common median reference. Some experiments also evaluated bipolar reference as an alternative method.

3.2.1.1. Common median reference - *by Philémon Roussel*

A each time step, the median value of all channels was computed. The resulting signal was subtracted from all channels to remove noise that was shared between all electrodes, such as line noise or electromagnetic interferences. Removing the median signal was found more robust to outliers than removing the average signal.

3.2.1.2. Bipolar reference

The common median reference is efficient to remove generalized noise over all electrodes, however it does not remove spatially localized noise. The bipolar reference method solves this issue by subtracting signals from adjacent electrodes. The signals of each pair of horizontally neighboring electrodes and each pair of vertically neighboring electrodes were subtracted from eachother. This creates new virtual channels representing the subtracted signals: on P5, the 72 channels were transformed into 63 horizontal features and 64 vertical features for a total of 127 channels.

3.2.2 Neural features

3.2.2.1. Spectral features

Spectrograms were computed from neural signals using a FFT with a moving hamming window of 200 ms, a 10 ms frame shift, and padding by symmetrizing the signal. The power spectral density of each frequency bands were averaged over 10 Hz bands from 0 to 200 Hz, resulting in 20 spectral features sampled at 100 Hz. In order to complete these features, the **slow cortical potentials** were extracted from neural signals by filtering between 0.5 and 5 Hz.

A total of 21 neural features were thus computed for each electrode signal of the ECoG dataset, which may cause a large memory usage for some decoding methods. For this reason, a **pooling of high gamma** bands was also considered: all 70 to 200 Hz bands were z-scored and averaged together, while the 60-70 Hz band was removed. This resulted in 8 neural features instead of 21.

3.2.2.2. Frontal and Temporal electrodes

P5 neural features were split into frontal and temporal categories. All features from electrodes placed above the lateral sylcus were considered as frontal, while the remaining features were considered as temporal (representation in Fig. 3.6). With these categories, 28 electrodes were categorized as frontal, and 44 as temporal (although some of those actually extend into the parietal lobe). The frontal electrodes covered the areas responsible for speech motor control, while the temporal electrodes covered the auditory cortex.

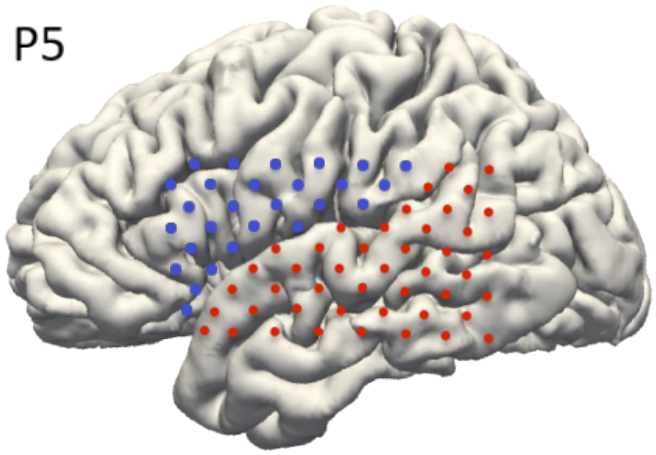


Fig. 3.6.: **Map of Frontal (blue) and Temporal (red) electrodes of P5 dataset.**

3.2.2.3. Phase features

Simon Julia, who was intern in the lab at the time, investigated neural features of speech based on the signals' phase. I then tested whether these features could be used to decode speech from neural activity. These features were solely computed on day 3 recordings of P5 for overt speech.

Neural signals were bandpassed in the 11 following frequency ranges: 0-0.5 Hz, 0.5-1 Hz, 1-2 Hz, 2-4 Hz, 4-8 Hz, 8-16 Hz, 16-32 Hz, 32-48 Hz, 52-64 Hz, 64-128

Hz, 128-256 Hz. The instantaneous phase of each bandpassed neural signals were computed using a Hilbert transform at 30 kHz. The phase difference between each pair of electrodes were then computed and downsampled to 1000 Hz, creating 2556 differences for each frequency band. Experiments showed that the variations of those phase differences were higher during resting state than during speech activity, specifically the variance of the derivative of phases differences was significantly lower for speech activity.

The **variance of the phase differences derivatives** were computed for each frequency band on 200 ms windows with a frame shift of 10 ms. The resulting 28 116 phase features (2556 differences * 11 frequency bands) were therefore sampled at 100 Hz, like spectral neural features. For the sake of conciseness, those features will now be referred to as *phase features*.

3.2.3 Acoustic contamination

During his PhD, Philémon Roussel showed that some ECoG recordings may be contaminated with acoustic signals. Some channels record speech sounds on top of neural activity, with varying intensity depending on the channel, the recording session, and the patient. A Matlab package was built by Philémon Roussel to assess if a dataset of neural electrophysiological data contains acoustic contamination and can be downloaded on zenodo (Roussel et al., 2021). I tested how such contamination could influence neural decoding of speech.

Both P2 and P5 were investigated and found to contain some acoustic contamination, although no recording bloc of P5 used in this manuscript were found to contain acoustic contamination. A statistical test showed that P2 contained contamination under 200 Hz (see Figure 3.7), as low as 100 Hz. Given that the spectral features of neural activity described in section 3.2.2.1 cover neural recordings from 0 to 200 Hz, they should be expected to be largely contaminated with P2's voice fundamental frequency that lies mostly between 100 and 160 Hz, while that of P5 lied between 200 and 300 Hz.

In order to safely remove influence of acoustic contamination on speech decoding, P2's spectral features were only considered up to 90 Hz, which was strictly lower that his voiced reached during the recordings. On the other hand, as both P5 and EC61 were women with higher pitched voice (over 200Hz) and no contamination was found in the recordings we used for this work, no spectral features were removed from those datasets.

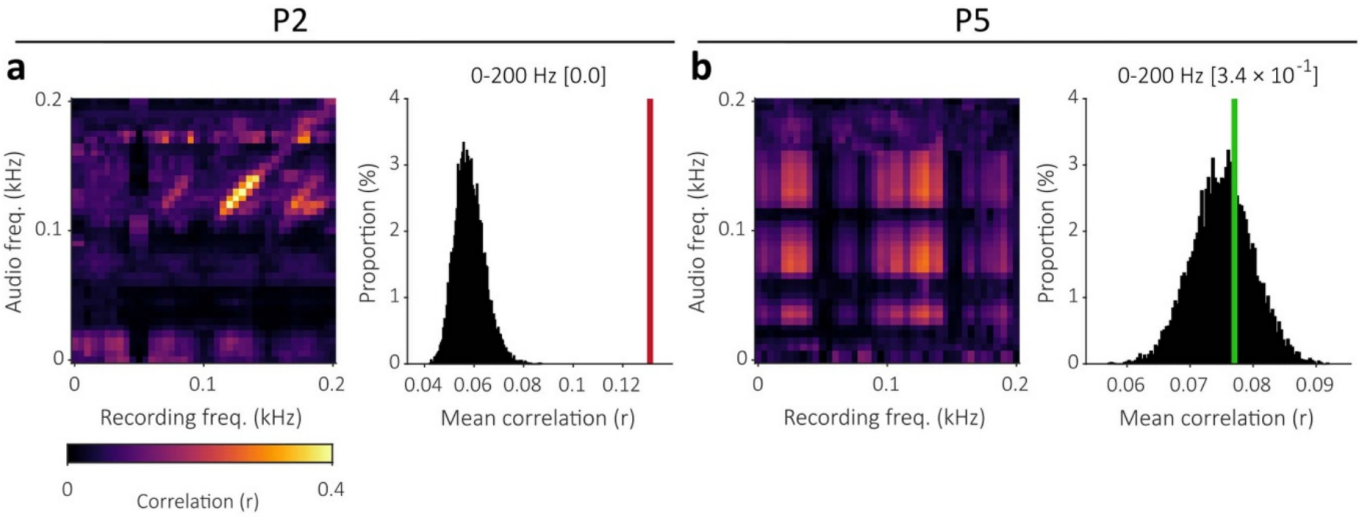


Fig. 3.7.: Contamination matrix between 0 and 200 Hz (left) and corresponding statistical assessment of contamination (right) for P2 and P5. The contamination matrices show correlations between audio recordings and ECoG recordings. The mean of the diagonal of the contamination matrix (vertical colored bar, red when statistically significant, green when not) is compared to the distribution of such values in 10 000 shuffled contamination matrices. The *p*-value corresponding to the estimated risk to wrongly consider the existence of contamination (P) is shown in square brackets for each dataset (P2 contaminated, P5 not contaminated).
from *Roussel et al., 2020*

3.2.4 Feature selection

The signals recorded by some ECoG electrodes might be too noisy or in brain areas unrelated to speech activity. In order to select which features described in section 3.2.2.1 were modulated by speech activity, a **Welch *t*-test with Bonferroni correction** was used to assess whether a feature had a significant change of activity between speaking and resting states.

Two brain datasets were built from each ECoG recordings, the first one containing only neural recordings during speech production, and the second one containing neural activity at rest only (*ie* no speech production or perception). The distribution of each feature's activity between rest and speech were compared using a Welch *t*-test, which returned one *p*-value per feature. A *p*-value lower than a given arbitrary threshold was interpreted as the two distributions being different, in other words that the corresponding feature did encode speech-related neural activity. Any feature that did not pass the corrected *p*-value threshold (0.05) was discarded. To account for multiple comparisons, a Bonferroni correction was applied to the *t*-test: for a target threshold of 0.05 and n the number of tests (equal to the number of features), the corrected threshold was set to $\frac{0.05}{n}$.

3.2.5 Context and delays

Actual speech production of the sound wave and its underlying cognitive processes are not synchronized. Indeed, the motor control of articulators requires planning and therefore happens before sound production, while the processing of auditory and somatosensory feedback happens after sound production. In order to take into account these cognitive processes that are not synchronized for speech decoding, two mechanisms were used: 1. a variable **time delay** between neural features and acoustic/articulatory features of speech, and 2. a variable **time context** that consisted in concatenating multiple consecutive frames of neural features to decode one frame of acoustic/articulatory features of speech.

3.3 Acoustic data processing

3.3.1 Preprocessing

Any DC offset was removed from audio sentences by subtracting their mean value from the signal. Resulting signals were then peak normalized and their average volume was set to -20 dB using automatic gain control in Matlab. Lastly, P2, P5 and EC61's sentences were resampled at 22050 Hz to mach BY2014's sampling rate using Matlab's *resample* function with the default antialiasing lowpass filter.

3.3.2 Source-filter representation

A mel cepstral and F0 analysis of speech was computed from audio recordings using SPTK (Imai, 2003). This source filter representation was motivated by the possibility for real-time synthesis of speech using an MLSA filter (Bocquelet et al., 2016c), which was also implemented by SPTK.

3.3.2.1. Mel cepstrum

Mel cepstrums of order 24 were extracted from audio recordings using SPTK. The signal analysis was performed with Blackman windows of 400 samples in input and 1024 in output with quadratic normalization, a frame shift of 220 samples and a frame length of 1024. The ε parameter was set to 10^{-4} to avoid errors in the

periodogram computations. The all-pass constant α described in equation 1.11 was set to 0.455 to accurately estimate the mel scale for a 22050 Hz sampling rate.

Due to the 220 sample frame shift, the resulting 25 mel cepstral coefficients were sampled at ~100.23 Hz. Each sentence was then resampled to 100 Hz in order to accurately match articulatory and neural features sampling rate. The resampling was performed by shape-preserving piecewise cubic interpolation of the signal with the *'pchip'* parameter of Matlab's *interp1* function.

3.3.2.2. F0

The F0 was extracted from the datasets using the SWIPE' algorithm from SPTK. The signal windowing used the exact same parameters used to extract mel cepstral coefficients described in section 3.3.2.1. After visual inspection of the dataset's spectrograms, the F0 search algorithm was constrained to 80-300 Hz for P2 and P5, and 120-330 Hz for EC61. The extracted F0 signals are equal to either the fundamental frequency when the signal is voiced, or 0 when it is unvoiced.

Like mel cepstral coefficients, the extracted F0 was resampled to 100 Hz to match articulatory and neural features. Due to the discontinuities in the F0 that had to be preserved, resampling was done with a nearest neighbor interpolation. F0 misdetections were filtered out by removing any F0 segment shorter than 50ms for P2 and P5, and 20ms for EC61. Those values were set after visual inspection of the data.

3.3.2.3. Synthesis

Speech audio was synthesized from mel cepstrum and F0 using SPTK's MLSA filter. The MLSA filter was excited by either a white noise source for voiceless signals, or an impulse train with a period changing according to the F0 for voiced signals. SPTK refers to this period as *pitch*, and requires it as the parameter controlling the generation of the excitation signal. A period of 0 defines by convention that no F0 is detected, and that the excitation signal should be white noise. Given a frame rate f_s (22050 Hz here), pitch was therefore reconstructed from F0 with the formula:

$$pitch = \begin{cases} \frac{f_s}{f_0} & \text{if } f_0 \neq 0 \\ 0 & \text{if } f_0 = 0 \end{cases} \tag{3.1}$$

Processing of the excitation signal and MLSA synthesis used the same α parameter and frame period used in the mel cepstral analysis presented in section section 3.3.2.1. Output waveforms were peak normalized and limited to avoid clipping, and loudness was set to -20 dB using automatic gain control.

3.3.3 Formants

Both formants analysis of P5 dataset and formant synthesis were implemented by Mohamed Baha Ben Ticha, a PhD student in the lab. I then tested decoding methods for prediction of speech formants from neural activity.

3.3.3.1. Formant extraction

Speech formants were extracted from P5 dataset using Python's Parselmouth package (Jadoul et al., 2018), which is a Python API for **PRAAT** (Boersma and Weenink, 2021). After visual inspection in PRAAT, a threshold for speech detection was set at -50 dB, so that formants were only computed in section of audio where the volume exceeded the threshold. After trying out values for formant analysis parameters on vowel sequences included in P5 dataset, the number of formants was set to 5, and the *'Formant ceiling'* was set to 5500 Hz. Only the first two formants **F1** and **F2** were kept as formant features, as those are mostly sufficient for characterizing vowels.

3.3.3.2. Formant synthesis

Formant synthesis was performed using **Klatt synthesizer** (Klatt, 1980). A C reimplementation of the original Fortran synthesizer (Klatt et al., 2015) was adapted into a real-time processing box of pulsIO, the lab's internal software for real-time processing of biosignals. The real-time implementation processes input parameters F0, F1 and F2 at 100 Hz, producing an audio buffer every 10 ms. As the synthesis controls were limited to the F0 and two first formants, only non nasalized vowels could be produced. The synthesizer however allows more control parameters, and could also produce nasalized vowels, fricatives and soft consonants with the adequate control parameters.

3.4 Articulatory data processing

3.4.1 Articulatory Data

Articulatory trajectories recorded by 3D Electromagnetic Articulography contains a lot of redundant information, as most of the trajectories can be characterized in the midsagittal plane, with some exceptions like lateral consonants which let air flow on the sides of the tongue (*e.g.* the english */l/*). MOCHA-TIMIT and PB2007 were already projected in the midsagittal plane, however that was not the case for BY2014. The 2 lips corners were removed from the 9 original sensors of BY2014 as they mostly move along the orthogonal axis. Then each sensor was projected on the midsagittal plane of the speaker using a PCA and keeping only the first two components. The resulting 14 articulatory features tracked the 2D trajectories of the upper and lower lips; tongue tip, back and dorsum; velum; and jaw.

In order to train articulatory-to-speech models, BY2014 and MOCHA-TIMIT were resampled to 200 Hz so that they match audio processing for this task (PB2007 was already recorded at 200 Hz, details in in section 3.5). In order to decode articulatory trajectories for each patients P2 and P5 from neural activity, those articulatory trajectories had to be inferred from BY2014 using the method described in section 3.4.2. For this use, and more generally all uses related to speech decoding, BY2014 was used at the original 100 Hz sampling rate which matches neural features.

3.4.2 Estimation of articulatory trajectories

Articulatory trajectories of P2 and P5 were estimated from the BY2014 dataset using Dynamic Time Warping (DTW) alignment computed on corresponding audio recordings.

3.4.2.1. Dynamic Time Warping

Dynamic Time Warping (DTW) is a standard algorithm used for measuring similarity of two temporal sequences without influence of tempo variations. It has been widely used for automatic speech recognition and speaker recognition (Sakoe and Chiba, 1978). Typically, the same sentence repeated twice by the same speaker should be very similar, while two different sentences spoken at the same speed should not be similar.

Given a sample-wise metric, DTW computes the non linear time distortion of two signals that minimizes the distance between them. The possible transformations include stretching or contracting the signals by duplicating samples. The optimal warping can be found by a dynamic programming algorithm that essentially maps one signal onto the other. Once signals are mapped onto eachother, the overall distance between them is not influenced by variation of speed. When two same sentences with different speeds and speakers are compared with DTW, a transformation that synchronizes the sentences is computed. This property was used to map BY2014 speech onto P2 and P5, who both read sentences from BY2014. This transformation was then applied onto the articulatory trajectories. The resulting estimated articulatory trajectories still represent BY2014's vocal tract, but as if it was speaking with the rate of P2 and P5.

The standard implementation requires both signals to have the same the number of samples, which was not the case as sentences of different speakers obviously have different durations. Signals were therefore resampled to have the same number of samples prior running the DTW. In order to not introduce side effects that would influence the DTW, signals were padded by symmetrizing their sides before resampling. A simple euclidean distance was used as a sample-wise metric for the dynamic programming algorithm.

3.4.2.2. Alignment features

Even though P2, P5 and BY2014 datasets contain the same sentences, they were spoken by different speakers of different genders. Moreover the EMA coils disturbed BY2014's speech, and so did the medical condition of P2 and P5 at the moment of the recordings. Thus, the tone, prosody, pitch and pronunciation of each speakers' recordings were different. Directly aligning raw waveforms of speech with DTW did not perform well. A slight improvement was obtained by aligning mel cepstrums instead, although without good success either.

In order to increase the accuracy of the alignment, I concatenated other features of speech to the mel cepstral coefficients (c_m), namely F0 and a boolean distinguishing speech and silent samples. Silence and speech were respectively labelled with 0 and 1 by an automatic speech detection algorithm based on audio envelope. F0 was set to 0 when no voicing was detected, which carries another boolean information about voicing. Like F0, detections of speech shorter than 50ms where filtered out to ensure that no background noises (mouse clicks, breath, movements) influence the

DTW. The resulting sample-wise distance between two signals S_1, S_2 can be written as:

$$\begin{aligned} d(S_1(t), S_2(t)) = &\sum_{m=0}^{24} \left(c_{1m}(t) - c_{2m}(t)\right)^2 \\ &+ \left(F0_1(t) - F0_2(t)\right)^2 \\ &+ \left(speech_1(t) - speech_2(t)\right)^2 \end{aligned}$$

Each feature can be weighted to ensure that no feature overly contributes to the euclidean distance and therefore to the alignment. A grid search was performed to test multiple weighting of each features and evaluate alignments with Pearson correlations. Best results were found by normalizing pitch by it's maximum value and normalizing all mel features by the absolute maximum value of the first mel (representing the power of the signal).

In this approach, the additional features act as soft constraints on the DTW that make sure that start and stop of voicing are aligned, as well as start and stop of speech. In between those boundaries, the mel cepstral coefficients become the main contributors to the alignment as they carry the most weight in the sample-wise distance. A visualization of the alignment of speech features and the resulting sound waves can be found in Fig. 3.8. The main advantage of soft constraints compared to hard constraints can be found in some situations where alignment features between two signals are not coherent. For example, in some P5 sentences vowels were voiceless, while they were normally voiced in BY2014. In those cases, voicing cannot be properly aligned, but the DTW proved to be robust and still properly aligned the mel cepstral coefficients.

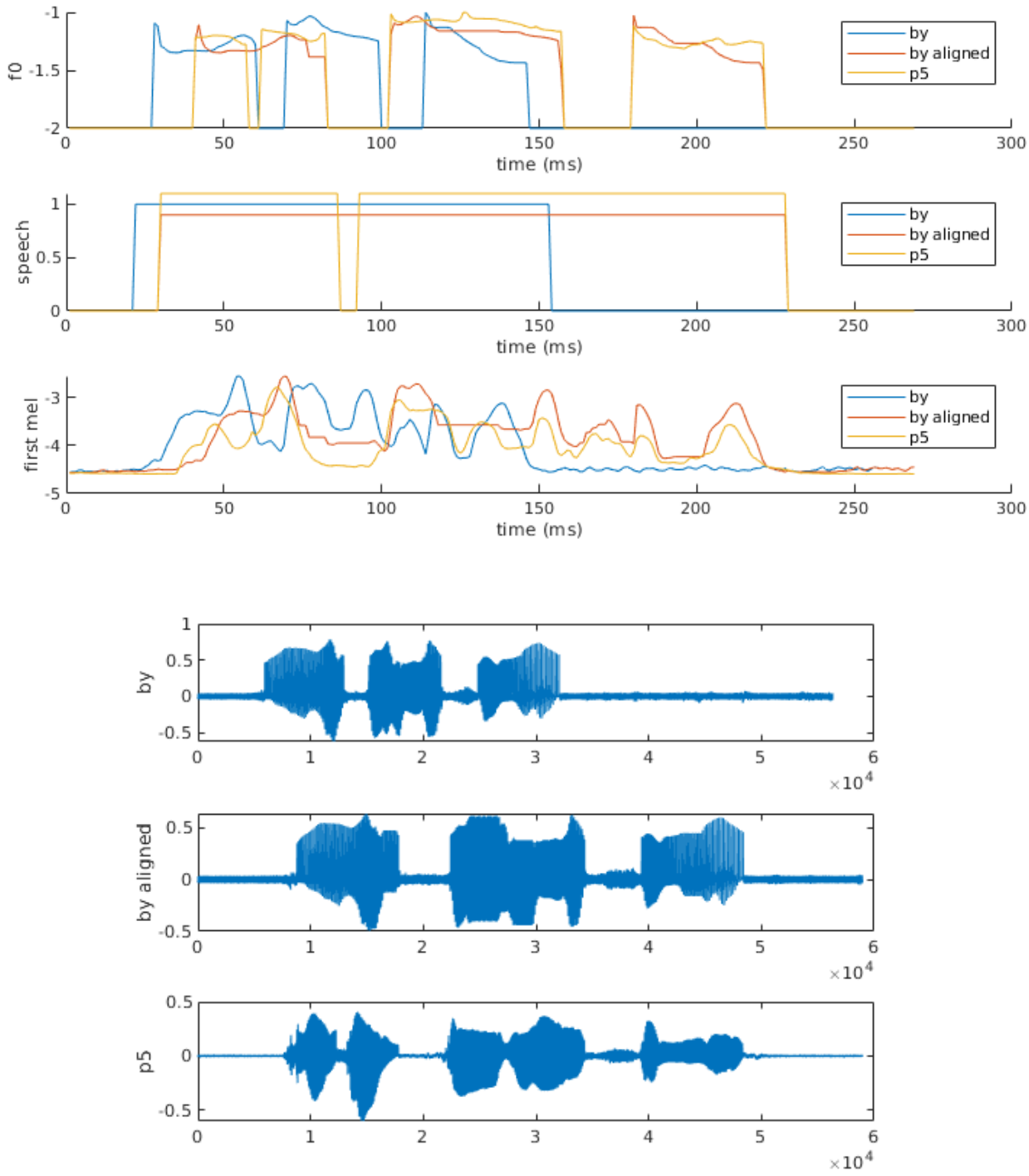


Fig. 3.8.: Alignment of the sentence "*Les affaires marchent bien*" from BY2014 onto P5. The top plot shows the alignment of the F0, speech and first mel. The bottom plot shows the resulting waveforms.
by: BY2014's sentence; *by aligned:* BY2014 sentence aligned on P5's matching sentence using DTW; *p5:* P5's sentence

3.5 Articulatory Synthesis

The work presented in this section has mostly been done under the direction of Thomas Hueber and Laurent Girin in Gipsa-Lab, Grenoble.

In order to reconstruct speech from decoded articulatory trajectories, a real-time-compatible articulatory-to-speech synthesizer was designed to predict mel features from articulatory features. Multiple neural networks inducing different latencies were compared with both objective and subjective evaluations.

3.5.1 Principle

As described in section 1.1, speech is physically produced by manipulation of airflow with the articulators. In other words, there is a direct causal relationship between the dynamic of the articulators and the produced speech. Articulatory synthesis consists in finding a transfer function that predicts speech sound from articulatory trajectories. The literature describes two main approaches to solve this problem: 1. physical models and 2. machine learning models.

In the **physical modeling** approach, the geometry of a standard vocal track is described in 1D,2D or 3D. Then a corresponding acoustic model of sound propagation computes a simulation of the sound wave going from the glottis to the mouth opening Birkholz et al., 2011. Physical models parameters can be finely controlled to produce on-demand speech with different voices but the synthesis quality is rather poor.

In the **machine learning** approach, a statistical model is trained to find a relationship between articulatory trajectories and acoustic features in a supervised fashion. Predicted acoustic features are then fed to a vocoder in order to synthesize the actual speech sound. Many different regression models have been used to map articulatory features into acoustic features with good performance: Gaussian Mixture Models (Toda et al., 2008), Hidden Markov Models (Hiroya and Honda, 2004) and neural networks (Kello and Plaut, 2004; Richmond, 2006). Best reported performance for real-time compatible models was achieved with deep neural network using temporal context (Aryal and Gutierrez-Osuna, 2016; Bocquelet et al., 2014). Overall state of the art performance was reported for bidirectional recurrent neural networks that are designed to handle long-term dependencies from both the past and the future (Liu et al., 2016; Taguchi and Kaburagi, 2018).

3.5.2 Regression methods

Aiming at designing a performant articulatory synthesizer, three neural network approaches were compared: 1. a simple feedforward Deep Neural Network (DNN), 2. DNNs with various added temporal contexts, and 3. a bidirectional Gated Recurrent Unit (GRU). Adding future temporal context to a DNN increases latency, while the biGRU effectively introduces arbitrarily long latencies by processing sentences as a whole.

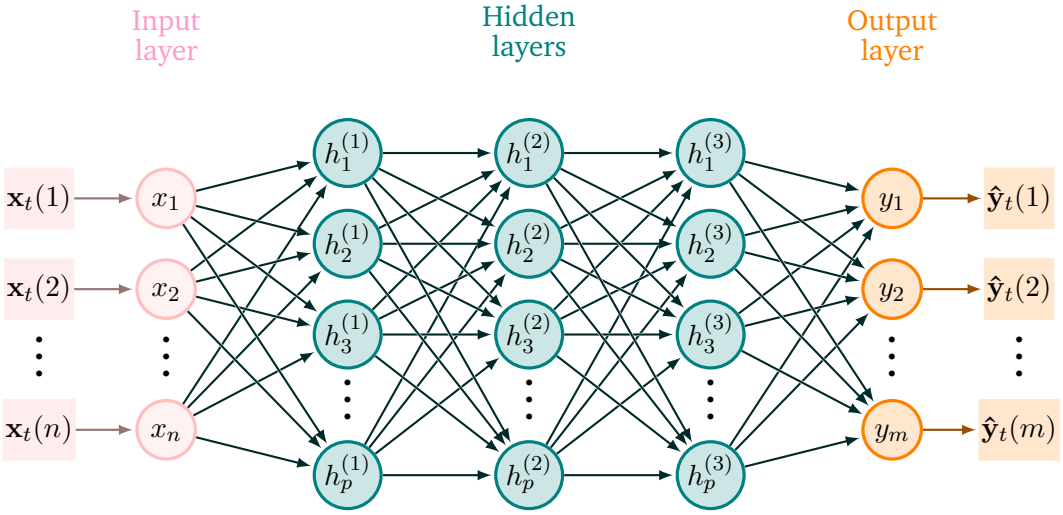


Fig. 3.9.: Feedforward Deep Neural Network. A DNN with 3 hidden layers is predicting a sample of m acoustic features $\hat{\mathbf{y}}_t$ from n articulatory features $\mathbf{x}_t$.
Circles represent individual neurons and squares represent individual input and output data features. Arrows represent the connections in between neurons and their associated weight.
Neurons x_i, $h_j^{(l)}$ and y_k respectively belong to the input, hidden and output layers. Input and output neurons are matching the number of features, while the number p of neurons per hidden layer can be arbitrary large.

3.5.2.1. Deep Neural Network

A feedforward **Deep Neural Network** (**DNN**), or **multilayer perceptron** consists of multiple interconnected layers of artificial neurons (Fig. 3.9). Layers are stacked so that each neuron of a layer is connected to every neurons of the next layer. Each neuron takes in input the weighted sum of the output of the previous layer's neurons and applies a non linear **activation function** to it. Let $\mathbf{x}_t$ a vector of articulatory features at time t and $\mathbf{y}_t$ a vector of the corresponding acoustic features at time t. On the input layer, each value of $\mathbf{x}_t$ is fed to the input of a dedicated neuron. Similarly

on the output layer, each value of $\mathbf{y}_t$ is predicted by the output of a dedicated neuron.

DNNs are trained to predict a sample $\mathbf{y}_t$ from $\mathbf{x}_t$ by adjusting all the weights of the network so that given and input $\mathbf{x}_t$, the predicted $\hat{\mathbf{y}}_t$ is as close to $\mathbf{y}_t$ as possible. The quality of the predictions are evaluated by a **loss function** $L(\hat{\mathbf{y}}_\mathbf{t}, \mathbf{y}_\mathbf{t})$ which is a metric of how close a prediction is from the expected output. When using only differentiable loss and activation functions, the network's weights can be optimized by backpropagation to minimize the loss function over a set of matching samples $\mathbf{x}_t$ and $\mathbf{y}_t$.

Training data is typically split into small **batches** of data samples and the update of the network's weights is computed after completing a batch. Once all the batches of the training set has been used to update the network's weights, an **epoch** is complete. Both the *batch size* and the *number of epochs* are hyperparameters that influence the performance of the network and the training duration.

3.5.2.2. Contextual DNN

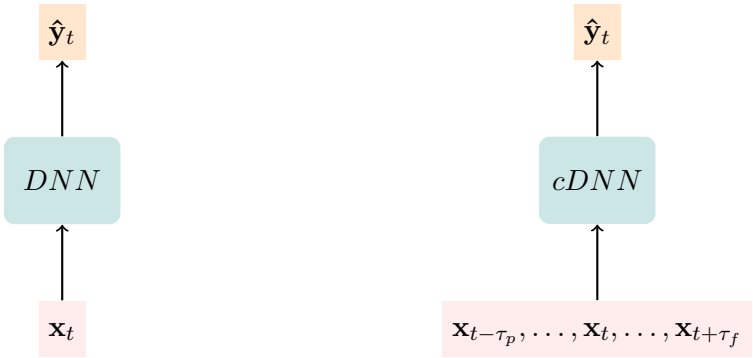


Fig. 3.10.: Contextual feedforward DNN. Left shows a simple DNN predicting a sample of acoustic features $\hat{\mathbf{y}}_t$ from a sample of articulatory features $\mathbf{x}_t$. Right shows a DNN with added temporal context (τ_p, τ_f): a single acoustic sample is predicted from multiple articulatory samples that span from $t - \tau_p$ to $t + \tau_f$.

Previous work investigated the effect of adding a past temporal context to the network input (Fig. 3.10), showing it is key to improve overall performance (Bocquelet et al., 2016c). Adding past context to a DNN consists simply in concatenating articulatory features from $\mathbf{x}_{t-\tau_p}$ to $\mathbf{x}_t$ into one vector to predict $\mathbf{y}_t$. This concept was extended to future temporal context, by also concatenating articulatory features from $\mathbf{x}_t$ to $\mathbf{x}_{t+\tau_f}$ to predict $\mathbf{y}_t$, which induces a latency of τ_f in the articulatory synthesizer.

Let f be the neural network's transfer function, and (τ_p, τ_f) its finite temporal context. The acoustic features at time t are estimated by:

$$\hat{\mathbf{y}}_t = f([\mathbf{x}_{t-\tau_p}^\intercal, \dots, \mathbf{x}_t^\intercal, \dots, \mathbf{x}_{t+\tau_f}^\intercal]^\intercal)$$

3.5.2.3. Bidirectional LSTM

While feedforward DNNs only handle finite temporal contexts, **Recurrent Neural Networks** (RNNs) are designed to process arbitrarily long sequences by managing an internal state: at each time step, the hidden layers not only process the new input data but also the output of the hidden layers at the previous time step, also called the hidden state $\mathbf{h}$. However basic RNNs fail when it comes to long term time dependencies, as their architecture does not allow the backpropagation's gradients to properly propagate through time. The **LSTM** (Long Short Term Memory) architecture was specifically designed to solve this issue by Hochreiter and Schmidthuber, 1997. It consists of multiple hidden layers responsible for managing the hidden state and the predictions.

The very best methods of articulatory to acoustic mapping reported in the literature are currently bidirectional LSTMs. The original LSTM handles arbitrary long dependencies from the past only, whereas **bidirectional LSTMs** also look at arbitrary long future dependencies (Graves and Schmidhuber, 2005). In practice, biLSTMs consist of one LSTM bloc that processes a sentence in one direction, and another LSTM bloc that processes it in the other direction. Let $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ the forward and backward hidden states at time t (see fig 3.11). Let f_y the transfer function of the layers responsible for prediction, $f_{\overrightarrow{h}}$ the transfer function of the layers responsible for managing the past hidden state, and $f_{\overleftarrow{h}}$ the transfer function of the layers managing future hidden state. Using the same notations as before, the sample acoustic features vector $\mathbf{y}_t$ at time t is predicted by:

$$\begin{aligned}
\hat{\mathbf{y}}_t &= f_y(\mathbf{x}_t, \overrightarrow{h}_{t-1}, \overleftarrow{h}_{t+1}) \\
\overrightarrow{h}_t &= f_{\overrightarrow{h}}(\mathbf{x}_{t-1}, \overrightarrow{h}_{t-2}) \\
\overleftarrow{h}_t &= f_{\overleftarrow{h}}(\mathbf{x}_{t-1}, \overleftarrow{h}_{t-2})
\end{aligned}$$

Out of the many variants of LSTMs, the **Gated Recurrent Unit** (GRU) architecture reduces the number of recurrent layers needed to handle the internal state, therefore

reducing the overall number of parameters for a similar network size. Although, GRUs don't have an explicit handling of memory, they perform comparably to LSTMs (Chung et al., 2014).

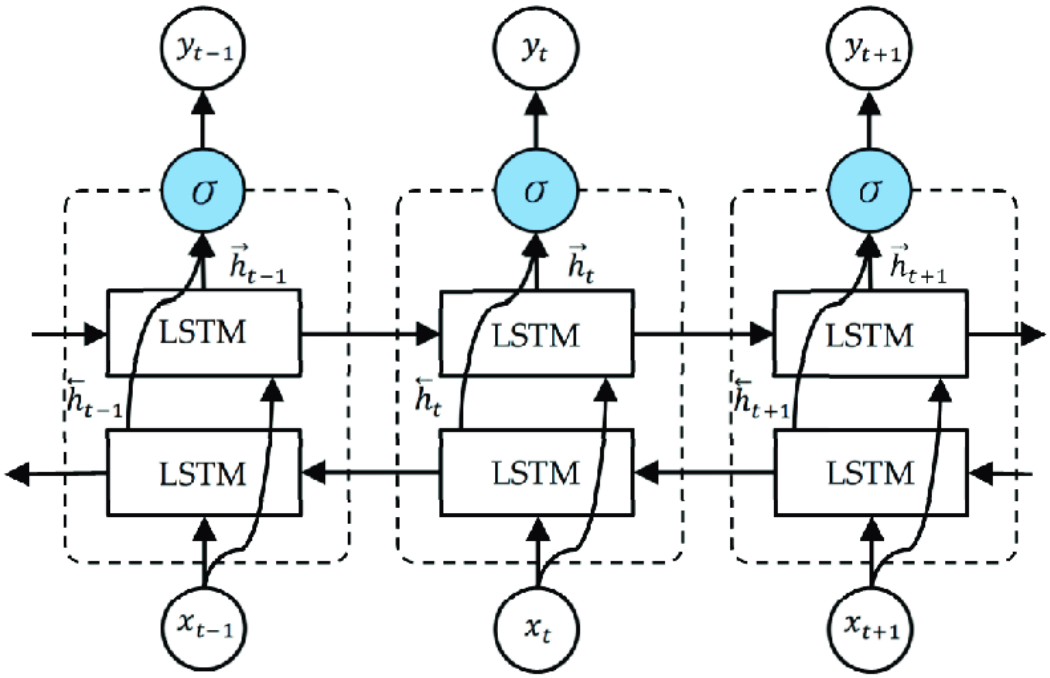


Fig. 3.11.: Bidirectional Recurrent Neural Network.
from *Li et al., 2020*

3.5.3 Training and evaluation

This section presents our framework for the training and objective evaluation of neural network models for articulatory-to-acoustic regression.

3.5.3.1. Training

Articulatory-to-acoustic neural networks models were trained to predict speech acoustic features from articulatory trajectories on individual EMA datasets. As already discussed in section 3.5.2.1, training a neural network actually consists in optimizing its weights so that the prediction error measured by a loss function (*loss* for short) decreases. Naively minimizing the loss on a training set of data does not usually achieve best performance on an unseen test set. Overly optimizing a model on a set of data leads to **overfitting**: the model learns statistical properties of the training data that do not generalize on the rest of the data.

In order to prevent overfitting, the model training and selection was automatically handled by **early stopping** (schematic in Fig 3.12). Data was split into a training set, a validation set and a testing set for final evaluation. After each epoch of tuning of the model's weights on the training set, the loss was computed on both the training set and the validation set. As training goes on, the training loss tends to decrease more and more, while the validation loss first decreases but then starts to increase again as the model overfits on the training set. Training was automatically stopped once the validation loss did not decrease for 20 epochs, and the model with the lowest validation loss was selected. The trained model was then evaluated on the testing set to assess its generalization power on unseen data.

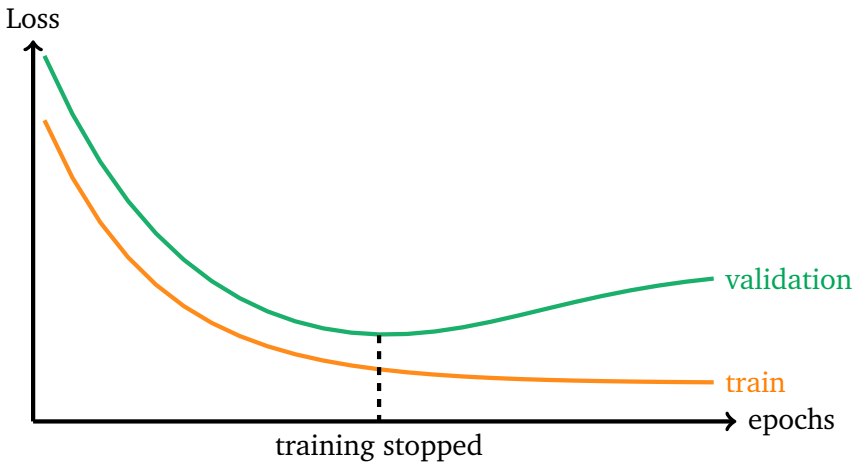


Fig. 3.12.: Early stopping. Training of neural networks is stopped when validation loss reaches a minimum, regardless of the training loss.

3.5.3.2. Crossvalidation

Articulatory-to-acoustic models were evaluated on a given dataset using **crossvalidation** (visualization in Fig. 3.13). Dataset's sentences were randomly split into k approximately equal folds (5 or 10 depending on our experiments). One of these folds was set aside for testing, one for validation and the rest was used to train the model. By permuting the folds and repeating the same procedure with a different test set, k models were trained and evaluated on each of the k folds. In the end, each sentence of the dataset was predicted by one of these models. Crossvalidation offers the double advantage of exploiting the dataset as a whole and evaluating the model design itself instead of a single neural network, as k instances of the model are evaluated on different data each time.

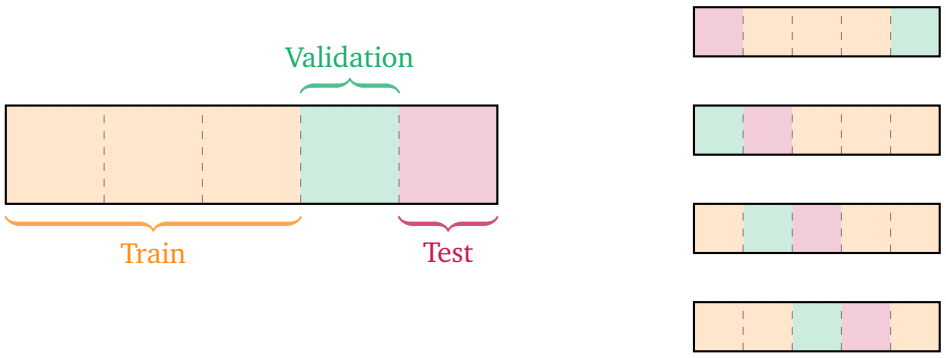


Fig. 3.13.: Example of a 5-fold permutation of data for crossvalidation. Test data is represented in red, validation in green and training data in orange. Each rectangle represents the same data randomly shuffled in the same order, split into 5 folds delimited with dashed lines. The fold used for testing is different for each permutation, so that all the data is used exactly once for evaluation. Training and validation sets are arbitrarily attributed to the other folds.

3.5.3.3. Evaluation metrics

The prediction quality of individual sentences were assessed by 3 different metrics. For a given sentence of T samples, let $Y_{t,m}$ be the matrix of its M acoustic features values and $\hat{Y}_{t,m}$ the matrix of the corresponding predicted features. The **mean squared error** (MSE) of the predicted sentence was defined as:

$$MSE(Y,\hat{Y}) = \frac{1}{MT}\sum_{t=0}^{T-1}\sum_{m=0}^{M-1}\left(Y_{t,m}-\hat{Y}_{t,m}\right)^2 \tag{3.2}$$

As mel cepstral coefficients were chosen as acoustic features, the sentences prediction can be measured in decibels by **mel cepstral distortion** (MCD):

$$MCD(Y,\hat{Y}) = \frac{10}{T\ln(10)}\sum_{t=0}^{T-1}\sqrt{2\sum_{m=0}^{M-1}\left(Y_{t,m}-\hat{Y}_{t,m}\right)^2} \tag{3.3}$$

Lastly, **Pearson correlation** of sentences and their predictions by the model were computed. Let $\mu_m, \hat{\mu}_m$ and $\sigma_m, \hat{\sigma}_m$ the mean and variance of the m^{th} acoustic feature of Y and $\hat{Y}$, respectively. The Pearson correlation of the m^{th} mel cepstral coefficient is defined by:

$$Corr(Y,\hat{Y})_m = \frac{1}{T-1}\sum_{t=0}^{T-1}\left(\frac{Y_{t,m}-\mu_m}{\sigma_m}\right)\left(\frac{\hat{y}_{t,m}-\hat{\mu}_m}{\hat{\sigma}_m}\right) \qquad ,\forall m \in [\![0, M-1]\!] \tag{3.4}$$

3.5.4 Subjective evaluation

While objective metrics provide some insight on the performance of an articulatory-to-acoustic synthesizer, they do not actually measure the quality of synthesized speech. All metrics presented in section 3.5.3.3 evaluate on an overall sentence how close predicted mel cepstral coefficients are from the actual ones. They measure predictions error without taking into account their impact on the MLSA synthesis. In order to accurately evaluate different articulatory-to-speech synthesizers, participants were asked to rate sentences predicted from multiple methods in a MUSHRA test.

3.5.4.1. MUSHRA test

Standing for **MUltiple Stimuli with Hidden Reference and Anchor**, **MUSHRA** is a method for subjective assessment of intermediate audio quality designed in the ITU recommendation BS.1534, 2014. MUSHRA tests compare a high quality **reference** signal with multiple matching **stimuli** with significant impairments. Stimuli are rated by the listeners on a scale from 0 to 100 divided in 5 categories: *'bad'* (0 to 20), *'poor'* (20 to 40), *'fair'* (40 to 60), *'good'* (60 to 80), *'excellent'* (80 to 100). All stimuli are rated in parallel with the instruction to accurately place them on the scale relatively to eachother. In order to calibrate the rating, the reference is hidden in the stimuli, as well as a **low anchor** which is a degraded stimulus that should be rated fairly lower than the evaluated stimuli.

The hidden anchor and reference also provide hindsight on the test. First, if the anchor were correctly designed, the legitimate stimuli should be rated from 20 to 80. Second, participants that fail to rate accurately the hidden reference should be screened out: exclusion threshold is set for rating references below 90 on more than 15% of the items.

The continuous rating design of the MUSHRA test is especially useful to compare stimuli with close overall quality. Its parallel rating provides the ability to performed paired *t*-tests, thus requiring only a few participants to obtain significant results. The recommendation states that no more than 20 participants should ever be necessary.

3.5.4.2. Experiment design

An online MUSHRA test adapted from Schoeffler et al., 2018 was setup to compare MLSA speech synthesis of multiple neural networks for articulatory-to-acoustic predictions.

Before starting the test, participants were asked to certify being native french speakers with no language or hearing issues. Participants were also required to use good quality headphones in a silent environment and to use an adequate volume (set on a test signal). The age and gender of the subjects was collected at the end of experiment, all data being anonymized and secured. Details on the experiment can be found in appendix A.2.

The test was split in two blocs for a total estimated length of 25 minutes. All the stimuli predicted by the different neural networks were part of the BY2014 corpus. In the first bloc, participants were presented with 9 **vowel-consonant-vowel** (**VCV**) sequences in a random order: *'aba'*, *'ada'*, *'aga'*, *'ibi'*, *'idi'*, *'igui'*, *'oubou'*, *'oudou'*, *'ougou'*. In BY2014, those were included into a contextual sentence: "*Tu t'appelles* **VCV** *c'est ça?*" ("*Your name is* **VCV***, right?*"). Participants were asked to rate the stimuli solely on the VCV sequence, not on the overall sentence quality. In the second bloc, participants were presented with 10 **short sentences** in a random order. The sentences were randomly selected prior to the experiment so that all participants were presented with the same stimuli. Participants were asked to rate the stimuli according to the overall quality of the sentences.

3.6 Speech Decoding

A source-filter representation of speech based on mel cepstrum and F0 was decoded from ECoG features by regression methods. Two different paradigms were investigated to decode mel cepstral coefficients: 1. direct decoding of mel cepstral coefficients using linear methods, and 2. decoding of articulatory trajectories using linear methods, an articulatory-to-acoustic neural network predicting the corresponding mel cepstral coefficients from articulatory trajectories. In order to provide a source signal for speech synthesis, the F0 was directly decoded from ECoG features using linear methods. Using the same methods, formants were also decoded to perform formant synthesis of vowels.

Fig. 3.14.: Online MUSHRA listening test. Screenshot of the web implementation of the MUSHRA test during the rating of an item. Reference and stimuli can be freely listened to by pushing the corresponding play button. The sliders are used to rate each stimulus compared to the reference.
accessible at (last visited 16/09/2021): *http://www.gipsa-lab.grenoble-inp.fr/~gael.legodais/tests_perceptifs/webMUSHRA_glegodais/?config=BY2014.yaml*

3.6.1 Neural features reduction

The number of neural features extracted from ECoG recordings for a single time frame may be very large, even after feature selection (section 3.2.4) that reduces the total number of neural features. In order to train a linear regression over neural data, the number of neural features is further reduced by PCA or PLS.

3.6.1.1. Principal component analysis

Principal Component Analysis (PCA) transforms the original features into a set of orthogonal features that are linear combinations of the original features. The algorithm of the PCA builds the orthogonal features by order of variance explanation of the data. The first PCA component is the linear combination of original data's features that explains the most variance in the data, the second component is build orthogonally to the first one to explain the most variance in the remaining data once removing the first component from it. The first PCA features therefore explain most variance of the data, and are likely more relevant to decoding of speech. Keeping only the first PCA components allows to greatly reduce the amount of features without loosing too much information. The number of PCA components to keep for decoding is an hyperparameter that should be optimized on a given dataset.

Let X a matrix of size (l, n) containing n neural features and l data samples, an efficient PCA analysis of X^T was implemented in Matlab using the **covariance method**. Before any computation, each feature column of the data matrix X was z-scored. The covariance matrix C of the resulting matrix was computed using Matlab's *cov* function, and its eigenvectors decomposition was computed using *eig* function. The eigenvectors were sorted by decreasing eigenvalue in a transformation matrix A of size (l, n) that defines a change of coordinate transforming a z-scored matrix X into its PCA components:

$$PCA(X^T) = X^T A \tag{3.5}$$

In order to reduce the number of PCA features to p components, only the first p of eigenvectors of X were kept from the PCA transformation matrix A, *i.e.* its first p columns.

3.6.1.2. Partial Least Squares

Partial Least Squares (PLS) transforms the features of an input data matrix X into a set of orthogonal features that minimize the covariance with a target matrix Y. Using the **SIMPLS** algorithm (Jong, 1993), the orthogonal PLS components are obtained as linear combination of the input data features. Much like PCA analysis, PLS transformation can reduce the amount of features without losing much information and the number of components is an hyperparameter that should be optimized for a given dataset. Unlike PCA however, the transformation takes into account a target dataset, which is especially suited for regression or classification tasks (Chao et al., 2010). More details about it can be found in section 3.6.2.3.

3.6.2 Linear decoders

Linear regression methods were trained to predict speech features from neural features. The following regression methods provide a transformation matrix M that can be applied to any neural data X with the correct representation to predict a matrix of speech features Y. This section describes the computation of such transformation matrices with different regularization and feature reduction methods.

3.6.2.1. Linear Regression

A simple linear regression was trained to predict a target matrix Y of size (l, m) with m speech features and l samples from an input matrix X of size (l, n) with n neural features. The resulting transformation can be summarized by a mapping matrix M and an error matrix ε such that:

$$Y_{i,j} = \varepsilon_{i,j} + M_{n+1,j} + \sum_{k=1}^{n} X_{i,k} \cdot M_{k,j}$$

The coefficients $M_{n+1,j}$ are called *bias*. Let X_b of size $(l, n+1)$ a matrix built by concatenating a $(n+1)^{th}$ feature filled with ones to X. The mapping matrix can then be computed by solving the following equation:

$$\widetilde{M} = \text{argmin}(||X_b \cdot M - Y||^2) \tag{3.6}$$

In practice, this was solved by computing the product of Y and the pseudo-inverse of X_b with Matlab's *pinv* function:

$$\widetilde{M} = pinv\,(X_b) \cdot Y \tag{3.7}$$

3.6.2.2. Ridge regression

Linear regressions are typically fitted on a training set of data and evaluated on a separate testing set. This can result in overfitting: when the regression performs well on the training set but does not generalize well on new data. In order to mitigate this issue, **ridge regression** introduces a form of regularization to linear regression called Tikhonov regularization. Using the same notations as for the simple Linear Regression, the transformation matrix $\widetilde{M}$ is defined as:

$$\widetilde{M} = \text{argmin}(||X_b \cdot M - Y||^2 + ||\lambda M||^2) \tag{3.8}$$

With λ a **regularization parameter** that needs to be evaluated. It is possible to find an analytical estimate of λ using a *L-curve criterion* (Calvetti et al., 2000). Let $\rho = ||X_b \cdot \widetilde{M}_\lambda - Y||^2$ and $\eta = ||\lambda \widetilde{M}_\lambda||^2$, ρ' and η' their derivatives. Considering $\widetilde{M}_\lambda$ and $\widetilde{Z}_\lambda$ such as:

$$\widetilde{M}_\lambda = \left(X_b^T X_b + \lambda^2 I\right)^{-1} X_b^T Y$$
$$\widetilde{Z}_\lambda = \left(X_b^T X_b + \lambda^2 I\right)^{-1} X_b^T \left(X_b \widetilde{M}_\lambda - Y\right)$$

we can show that:

$$\eta' = diag\left(\frac{4}{\lambda} \widetilde{M}_\lambda \widetilde{Z}_\lambda\right)$$
$$\rho' = -\lambda^2 \eta'$$

Plotting ρ against η for different values of λ typically yields a L-shaped curve. Using the previous equations, we can show that the curvature K of the L-curve is:

$$K = \frac{2\lambda}{(1 + \lambda^4)^{\frac{3}{2}} |\eta'|} \tag{3.9}$$

A good estimate of λ is found at the L corner of the curve, where the curvature K is maximal. The λ value for which the curvature is maximal can be estimated by grid search.

Another approach to estimate λ would be to try out different values of lambda in a grid search and evaluate the best value with *crossvalidation*. Both methods share similar computing complexity and memory load which increases quickly with the number of input features of X.

Lastly, it is also possible to compute different values λ_i for each output feature $\mathbf{y}_i$. Let $\mathbf{m}_i = (M_{i,1}, M_{i,2}, \cdots, M_{i,m})$, without increasing the computational cost, $\Lambda = (\lambda_1, \lambda_2, \cdot, \lambda_n)$ can be found by solving the following optimization problem:

$$\forall i \in [\![1, n+1]\!], \quad \widetilde{\mathbf{m}}_i = \text{argmin}(||X_b \cdot \mathbf{m}_i - Y_i||^2 + ||\lambda_i \mathbf{m}_i||^2) \tag{3.10}$$

As for a simple linear regression, this problem was solved by computing inverse matrices using Matlab's backslash operator \ :

$$\forall i \in [\![1, n+1]\!], \quad \widetilde{\mathbf{m}}_i = \left(X_b^T X_b + \lambda_i I_{n+1}\right) \cdot \left(X_b^T \cdot \mathbf{y}_i\right)^{-1} \tag{3.11}$$

3.6.2.3. Partial Least Squares Regression

Using the same notations, a **Partial Least Squares** transformation of both X and Y can be computed prior to a simple linear regression (see section 3.6.1.2). For a PLS regression with p components, a matrix decomposition of the input and target data is computed as follows:

$$X = T \cdot P^T + E$$
$$Y = U \cdot Q^T + F$$

Where T, U are of size (l, p), P is of size n, p and Q of size m, p and matrices E and F are the error terms such that $cov(T, U)$ is maximized. Given a transformation matrix M_p of size (p, p) such that:

$$M_p = \text{argmin}(||T \cdot M_p - U||^2)$$

The mapping matrix M that maximizes covariance between input data X and target data Y can be estimated as $M = P \cdot M_p \cdot Q^T$. In practice, the whole PLS transformation and regression was performed by Matlab's built-in function *plsregress* that uses the SIMPLS algorithm (Jong, 1993) to estimate M.

3.6.3 Decoding paradigms

3.6.3.1. Direct decoding

Mel cepstral coefficients and F0 were directly predicted from neural features by training multiple linear methods described in section 3.6.2. Prior to training regressions, input and target data were preprocessed regardless of the nature of the neural and acoustic features.

First, input and target features were z-scored. Second, a time delay was optionally applied between input and target data. Third, an PCA decomposition was optionally performed to reduce the number of neural features. Finally, various amounts of temporal context was added to neural features by concatenating past and future frames of neural activity to predict a single frame of acoustic features; the same technique was already described in section 3.5.2.2.

Linear regressions, with or without regularization were optionally combined with a PCA decomposition as described in the previous paragraph, while PLS by design performs feature reduction altogether with regression after any other preprocessing.

3.6.3.2. Indirect decoding

Using the same linear decoding methods, articulatory features of speech were also decoded from neural activity. Mel cepstral coefficients could then be predicted from decoded articulatory trajectories by articulatory-to-acoustic neural networks trained on BY2014, as described in section 3.5.2.

The articulatory-to-acoustic neural networks were trained to predict BY2014's mel cepstral coefficients from BY2014's articulatory trajectories. Although P2 and P5 datasets articulatory trajectories were estimated from BY2014, their temporal structure was different. Neural models were therefore **finetuned** to better fit the patient's data: the network's weights after training on BY2014 were used as initialization weights for training the model to predict the patient's mel cepstral coefficients from its decoded articulatory trajectories. With exception of the neural network initialization using a pretrained model, the training method is exactly the same as the one described in section 3.5.3.1. With finetuning, the articulatory-to-acoustic neural model predicts the patient's mel cepstrum instead of BY2014's.

3.6.3.3. Formants and F0 decoding

Both the voice's fundamental frequency F0 and the two first formants F1 and F2 were also decoded from neural activity using linear methods, following the methods described in section 3.6.3.1. While these methods are well suited to predict continuous signals like mel cepstral coefficients or articulatory trajectories, they cannot properly reproduce discontinuities such as the ones induced by voicing. Formants and F0 are indeed tracking a continuous value during voiced speech segments, but are set to 0 during voiceless speech segments.

As a first step towards improving the decoding of formants and F0, a second linear regression was trained to predict formants and F0 on voiced segments only. The trained linear decoder was then applied on complete sentences, predicting a non zero value at each time sample. The predicted F0 and formants were then multiplied by a **gate** classifying voiced and voiceless segments: the gate's value is 0 for voiceless frames and 1 for voiced frames. This method can correctly reproduce discontinuities of F0, F1 and F2 induced by the voicing.

A boolean index of voiced segments extracted from original datasets was used as a theoretically optimal gate. Linear models were trained to predict formants and F0 on voiced data only and then applied to decoding of complete sentences. The decoded formants and F0 were then gated by the extracted boolean index, as if a perfect gate was decoded from neural activity.

3.6.4 Speech synthesis

Speech was synthesized from decoded acoustic features either from source-filter representation or from speech formants.

3.6.4.1. Source-filter synthesis

Speech was synthesized from the decoded mel cepstral coefficients and F0 using MLSA synthesis (see sec. 1.2.6.1) as described in section 3.3.2.3. While the decoded mel cepstral coefficients did not require any processing before synthesis, the MLSA filter required to compute the *pitch* from F0. When the F0 was predicted using a voicing gate, the equation 3.1 could directly be used to compute the *pitch*. On the other hand when the F0 was predicted by a regression only, any F0 value below 80 Hz was set to 0. As regression methods do not properly reproduce discontinuities,

the predicted F0 sometimes takes low or negative values. This should not normally happen as the voice fundamental frequency rarely reaches values lower than 100 Hz for humans. Due to the inverse in the pitch formula, low F0 values induced very high pitch values which produced audio artefacts in the synthesis.

3.6.4.2. Formant synthesis

Vowels were synthesized from F0, F1 and F2 using Klatt synthesizer, as described in section 3.3.3.2. As the regression methods used for decoding were not constrained to positive values, all negative values of F1 and F2 were set to 0 before synthesis. For the same reason, F0 values below 80 Hz were also set to 0 in order to distinguish voiced and voiceless segments.

3.6.5 Evaluation framework

Decoding methods were evaluated by comparing the speech features predicted from brain activity with the true features extracted from the patient's speech. The decoding models were evaluated on all the data using a 10-fold crossvalidation.

3.6.5.1. Crossvalidation

Every decoding method (linear and DNN) were evaluated on a **10-fold crossvalidation**. Each corpus was randomly split in 10 folds that contain approximatively the same number of sentences. One fold constituted the testing set, the others constituted the training set. Each fold was used for testing once following the principle of section 3.5.3.2 until the model has been evaluated on all the corpus.

In case of the indirect decoding paradigm, the linear model trained on the 9 train folds predicted the articulatory trajectories of all 10 folds. An articulatory-to-acoustic DNN was then finetuned on the predicted articulatory trajectories of the same 9 train folds and evaluated on the 10^{th} fold.

During each fold of the crossvalidation, data was preprocessed in specific (optional) steps listed here in chronological order:

1. Normalization of input and target features by z-score (computed on training set)
2. Shift in time of the input with respect to the target

3. Feature reduction by PCA
4. Concatenation of input frames to create a time context

Normalization and PCA transformations were computed on the training set and then applied to both training and testing sets. It is to be noted that the feature selection was already computed on the complete datasets before the crossvalidation.

3.6.5.2. Evaluation of predicted speech

Decoding methods were evaluated by comparing predicted sentences with ground truth sentences using Pearson correlation and mean squared error (see section 3.5.3.3). Thanks to the crossvalidation, the predictions of all sentences of the datasets could be evaluated.

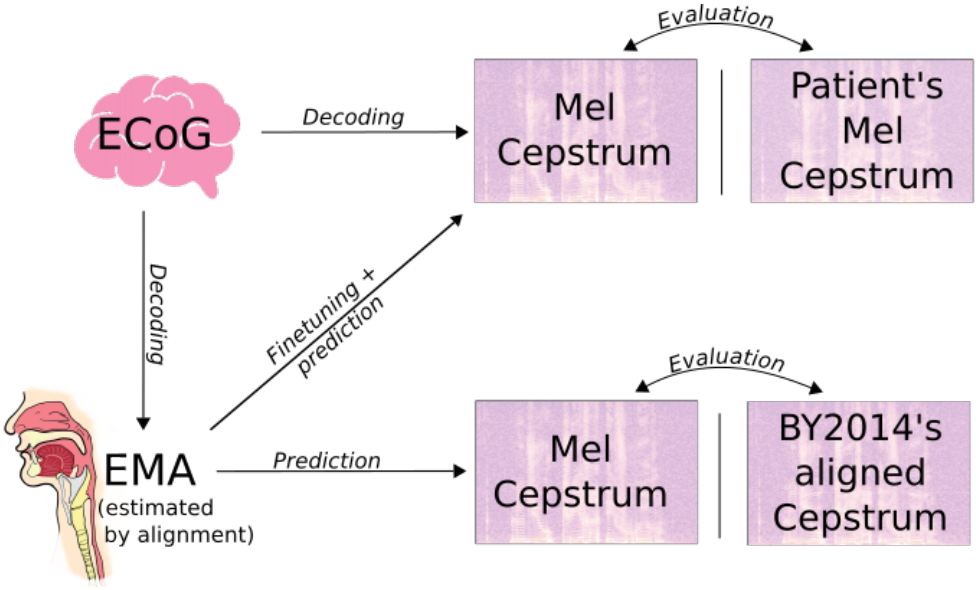


Fig. 3.15.: Evaluation of speech decoding from neural activity. Mel cepstral coefficients are decoded either 1. directly from neural activity or 2. from articulatory trajectories (EMA) decoded from neural activity. Mel cepstral coefficients directly decoded from neural activity or predicted from decoded articulatory trajectories with finetuning are evaluated against the patient's original mel cepstrum. Mel cepstral coefficients predicted from decoded articulatory trajectories without finetuning are evaluated against BY2014's mel cepstrum aligned on the patient's mel cepstrum using DTW.

Features predicted from direct decoding methods were compared to the patient's ground truth features. Mel cepstral coefficients predicted from articulatory methods with finetuning were also compared with the patient's mel cepstral coefficients, while mel features predicted without finetuning were compared to BY2014's mel cepstral coefficients aligned on the patient's features by DTW (Fig. 3.15).

Formants and F0 were either evaluated on complete sentences or on voiced segments only to isolate the contribution of voicing prediction. The voiced segments selected for evaluation are the voiced segment of the ground truth extracted features. Finally the accuracy of the linear classifier predicting the voicing was evaluated on complete sentences.

4 Results

4.1 Neural features modulated by speech production

Speech decoding from neural activity requires to compute a set of features that represent the relevant neural activity. Before decoding speech from neural activity, each dataset's neural features were therefore computed as 21 frequency bands for every channels (see section 3.2.2.1). We'll refer to an individual frequency band of an individual channel as a **feature**. This section investigates neural features with respect to speech production.

4.1.1 Feature selection

Spectral features of neural activity may represent some irrelevant information such as noise, which may lead to overfitting. In order to prevent overfitting, only relevant features of a given dataset should be selected before decoding. Therefore, for a given dataset, each feature was automatically selected by Welch's *t*-test with a Bonferroni correction (see section 3.2.4). This method ensures that only neural features that are modulated by speech production are used for speech decoding.

4.1.1.1. Selected features

On P2 dataset, only 2687 features out 5376 (21 frequency bands, 256 channels) were left after the statistical feature selection. While beta bands from 20Hz to 40Hz and high gamma bands from 60Hz to 140Hz showed overall lower p-values and higher selection rate, the statistical selection mostly removed entire channels. Fig. 4.1 shows that channels that were almost entirely discarded are concentrated in a large section of the grid, from channel 126 to channel 230. It is a known issue that many channels of P2 did not properly record neural activity. Due to its high electrode density, the ECoG array used for the recording was too rigid to closely fit the surface of the brain. A previous visual inspection of the recordings by Bocquelet, 2017 showed noise or no signal at all in 125 out of 256 electrodes, largely overlapping statistically selected electrodes.

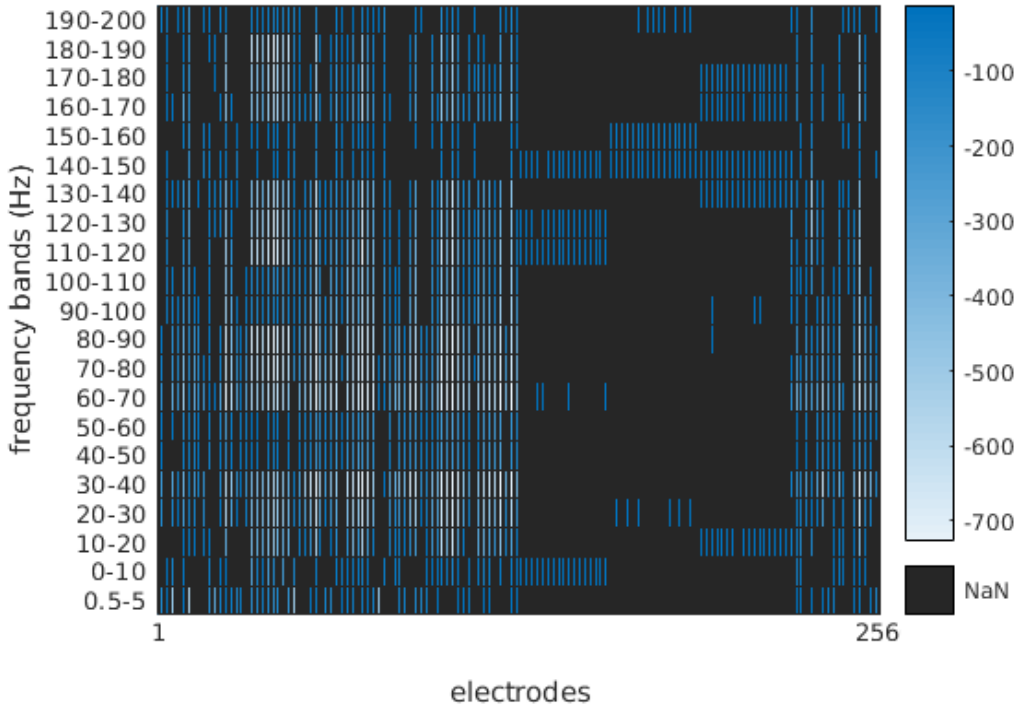
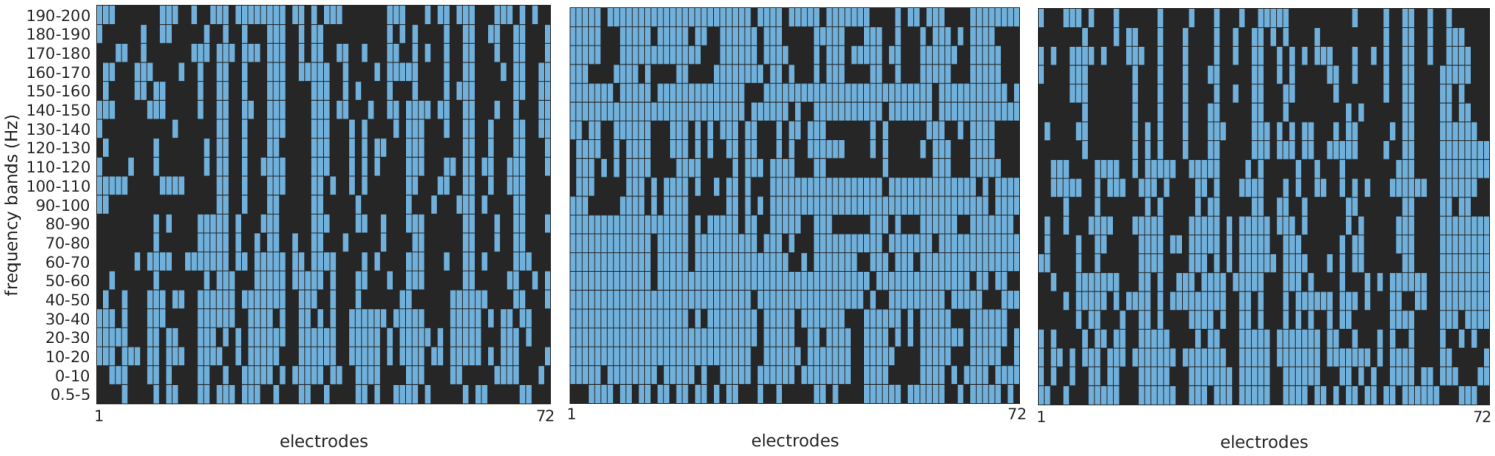


Fig. 4.1.: Feature selection on P2.
Each cell shows the Bonferroni corrected Welch p-value of a feature made of a frequency band (rows) in a channel (columns). The color scale represents the magnitude of the p-value in power of 10. Discarded features ($p > 0.05$) are colored in black.



(a) D1 (b) D2 (c) D3

Fig. 4.2.: Feature selection on day 1,2 and 3 of P5. Discarded features ($p > 0.05$) are colored in black, while selected features are colored in blue.

	P2	P5 day 1	P5 day 2	P5 day 3	P5 days 1,2,3
selected	2687	669	1114	702	1455
total	5376	1512	1512	1512	1512

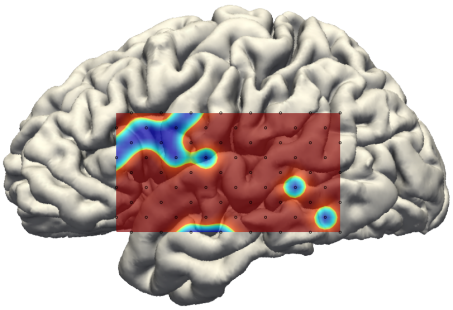
Fig. 4.3.: Number of selected features. The first row displays the number of statistically selected features using Welch's *t*-test for each dataset, compared to the total number of features in the dataset.

On P5 dataset, 669 and 702 features out of 1512 were selected on day 1 and day 3 of the recordings, while 1114 out of 1512 features were selected on day 2. When grouping the three separate days of recording however, almost all the features were selected.

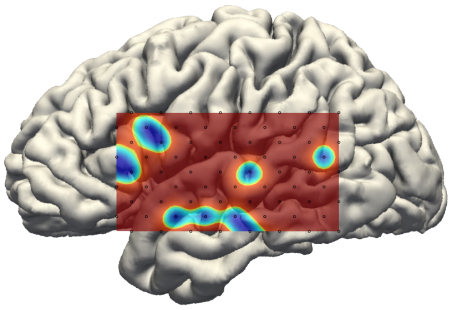
4.1.2 Mapping

The Welch *t*-test between rest and speech activity used for feature selection provided *p-values* that were mapped onto a brain model of P5 using NeuroMap software (Abdoun et al., 2011). Fig. 4.3 shows a map of *p-values* in each neural features frequency bands computed from the reading condition of day 2 of P5. Like for feature selection, *p-values* above 0.05 with Bonferroni correction were considered to indicate no speech related activity and are colored in blues on the mapping. The *p-values* below 0.05 were represented on a linear scale, with values closer to 0 in dark red and values close to 0.05 in orange. NeuroMap shows the accurate *p-value* associated with a given electrode and interpolates in between electrodes to obtain a complete map over the area of the ECoG array. Interpolated values were represented in light blue to green for values close to 0.05. A safe way to read the mapping in Fig. 4.3 would be to consider that blue colored electrodes likely did not record much speech activity in a given frequency band, while orange/red electrodes did. Anything in between electrodes is a guess based on interpolation, which may create some artifacts and makes it difficult to read light blue to orange areas.

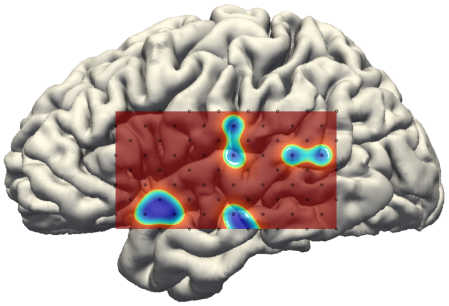
Neural activity in Broca's area seems to be modulated by speech activity in the frequency ranges between 10-30Hz, 40-50Hz, 60-80Hz, 100-140Hz and 190-200Hz. Neural activity in the motor cortex is only partly covered by the ECoG grid and seems to be modulated mostly below 30Hz, and also in the frequency ranges from 40 to 70Hz and from 100 to 140Hz, while it remains partly modulated in most frequency ranges. Speech related activity seems to be recorded across most if not all the entire superior temporal gyrus below 10Hz, from 60 to 80Hz and from 130 to 150Hz. In other frequency bands, electrodes of the superior temporal gyrus that did not show speech activity were mostly located in its anterior and parietotemporal parts.



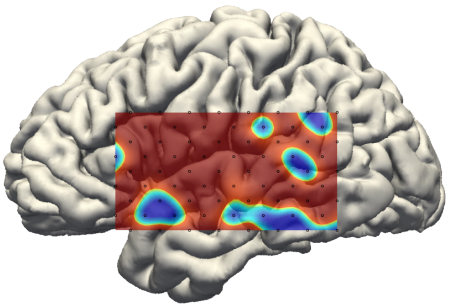
(a) Slow Potentials: 0.5-5 Hz



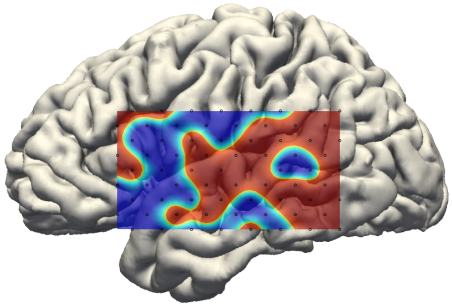
(b) 0-10 Hz



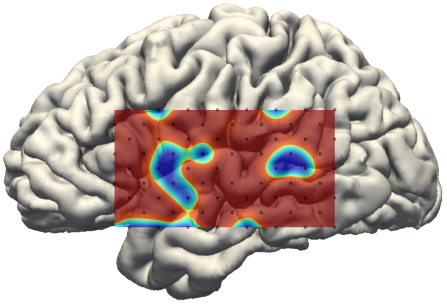
(c) 10-20 Hz



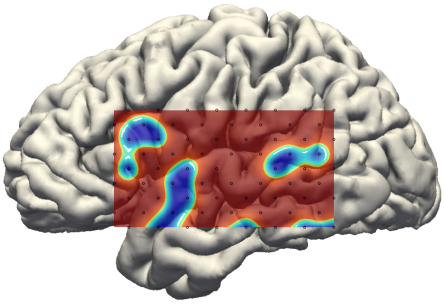
(d) 20-30 Hz



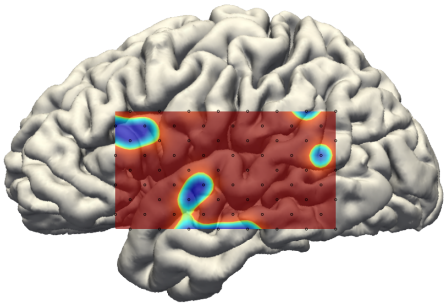
(e) 30-40 Hz



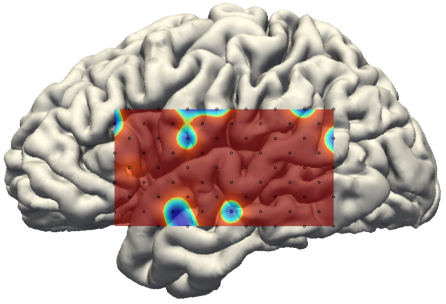
(f) 40-50 Hz



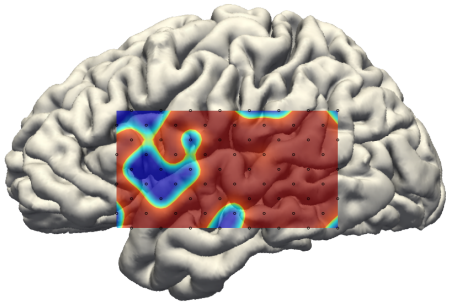
(g) 50-60 Hz



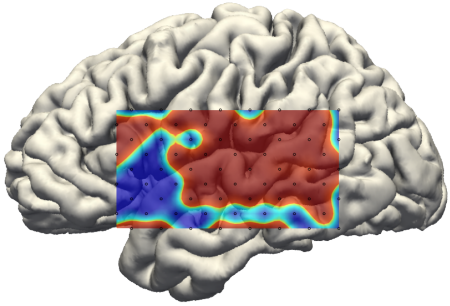
(h) 60-70 Hz



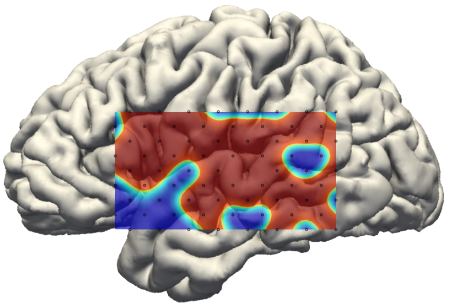
(i) 70-80 Hz



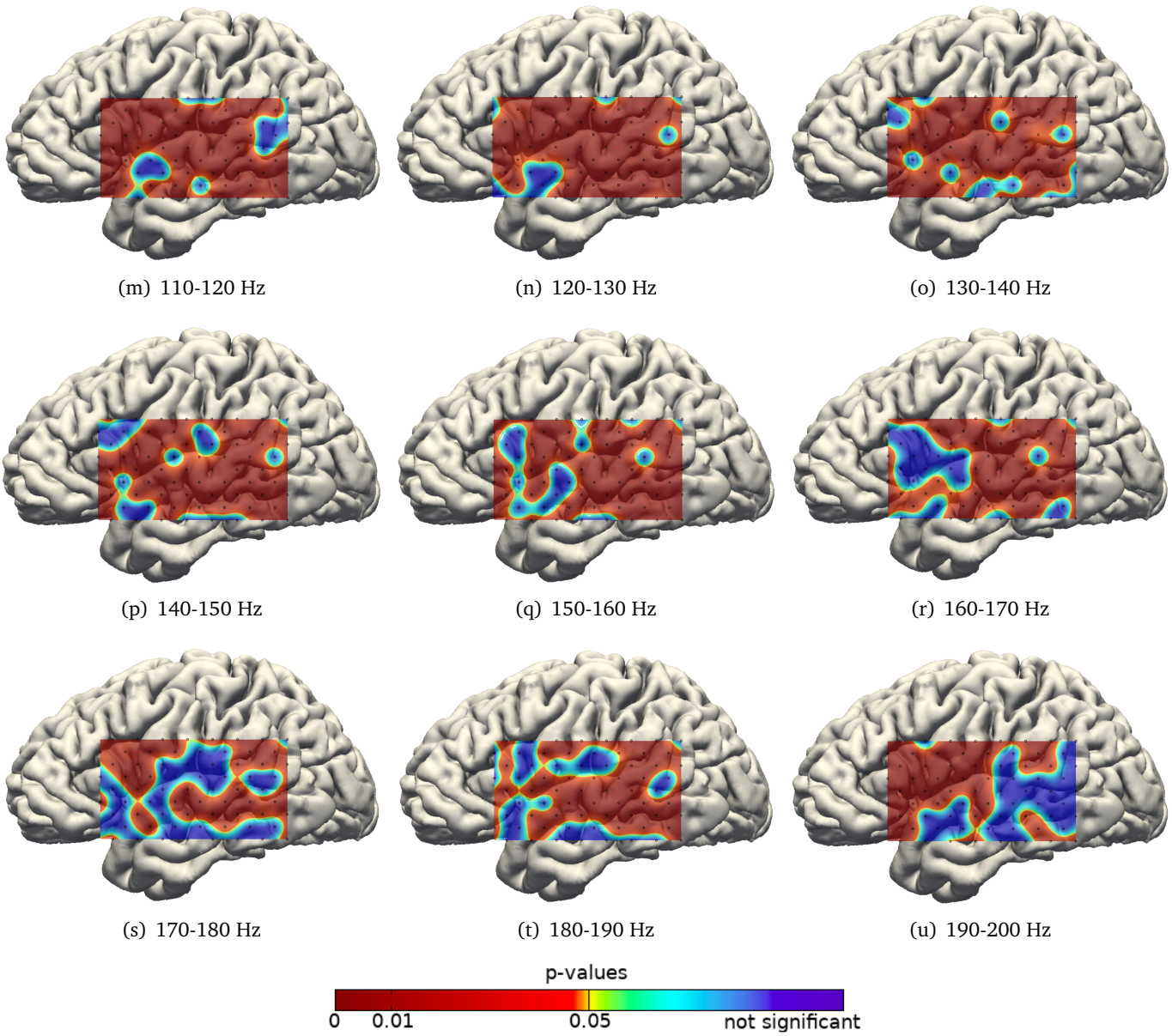
(j) 80-90 Hz



(k) 90-100 Hz



(l) 100-110 Hz



(m) 110-120 Hz (n) 120-130 Hz (o) 130-140 Hz

(p) 140-150 Hz (q) 150-160 Hz (r) 160-170 Hz

(s) 170-180 Hz (t) 180-190 Hz (u) 190-200 Hz

Fig. 4.3.: Mapping of feature selection p-values on P5 brain model for every frequency bands.
Red to orange show lower to higher p-values, up to 0.05 (with Bonferroni correction) on a linear scale. All electrodes with non significant p-values (>0.05 with Bonferroni correction) are shown with deep blue at the right end of the scale. Green to light blue values are interpolations computed by NeuroMap software.

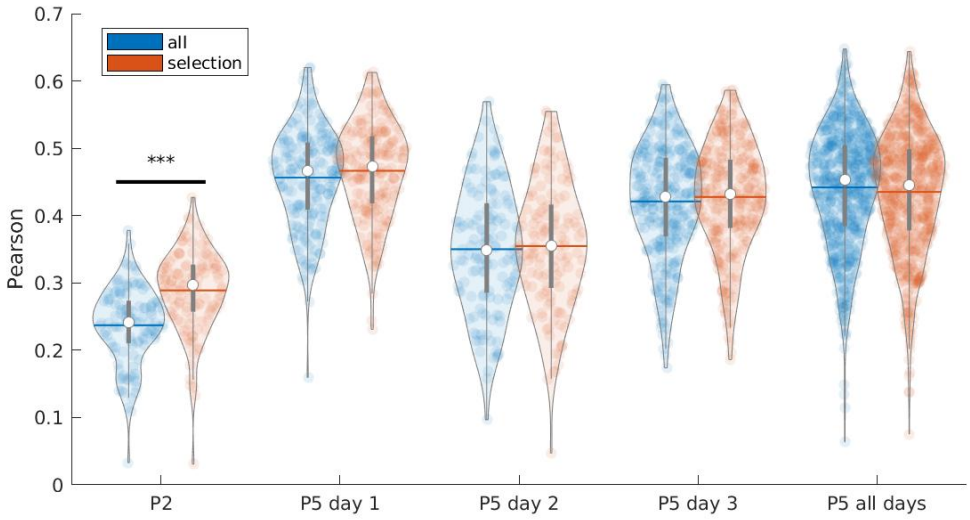


Fig. 4.4.: Effect of feature selection on linear decoding of speech. Mel cepstral coefficients were decoded from neural activity using linear regression and a PCA reduction to 100 components. Decoding from all neural features (blue) was compared to decoding from selected neural features (red). PCA reduction was computed after feature selection. Each point in violin plots represents the average correlation of mel cesptral coefficients for one sentence of the dataset. The white dot shows the median sentence and the colored horizontal line shows the mean correlation. Statistical significance was assessed by a Wilcoxon signed-rank test. ***: $p < 0.001$

4.2 Direct decoding of acoustic speech features using linear methods

This section focuses on decoding of speech from P5 dataset using linear methods. All models were trained on read and repeat speech conditions of day 1,2 and 3 of P5, with feature selection. Decoding was evaluated over 10-fold crossvalidation where 90% of the data was used for training and the remaining 10% for evaluation.

4.2.1 Influence of feature selection on linear decoding

A linear model was trained to predict mel cesptral coefficients from 1. all neural features and 2. selected neural features on P2, individual days of P5 and the complete P5 dataset. For each dataset, the model was evaluated in a 10-fold crossvalidation, effectively predicting all sentences of the dataset by 10 different models. In order to evaluate the effect of feature selection on speech decoding, Figure 4.4 shows

for each dataset the mean correlation of predicted mel cesptral coefficients with or without feature selection. Whether there was some feature selection or not, a PCA was computed on the remaining neural features and only the first 100 components were used for decoding, then a past and future context of 10 frames were used for decoding.

Although less than half of the neural features were selected on day 1 and 3 of P5, the decoded mel cesptral coefficients mean correlations were very similar when using all or selected neural features. This was also the case for day 2 of P5 where more features were selected, as well as for the complete dataset for which almost all features were selected. On the other hand, feature selection significantly improved the decoding of speech from P2 dataset (Wilcoxon signed rank test: $p = 6.5e^{-16}$), with a mean correlation increasing from 0.24 to 0.29.

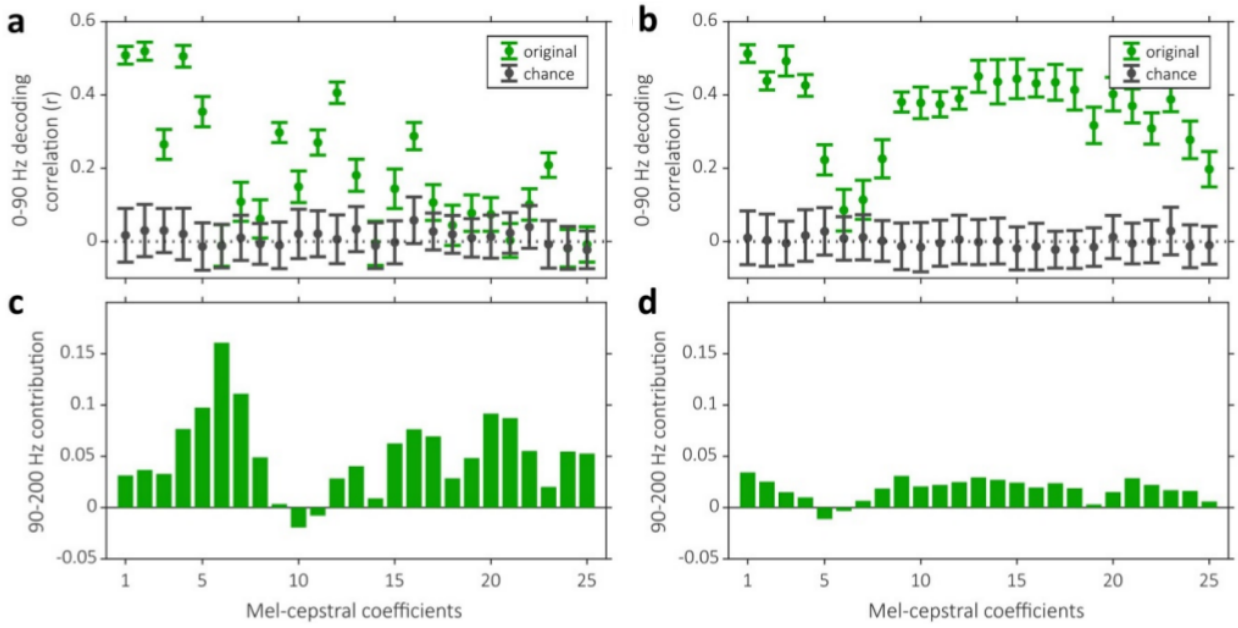


Fig. 4.5.: Influence of frequency bands with acoustic contamination on speech decoding. Left column shows decoding results for P2 (read condition), which includes acoustic contamination in the 90-200Hz frequency range. Right column shows speech decoding from P5 (day 3, read condition only). (a-b) Correlations of decoded mel cepstral coefficients from neural spectrograms between 0 and 90Hz (green) and matching chance levels computed by shuffling neural signals (grey). (c-d) Relative change of performance when adding 90-200Hz features. See Roussel et al., 2020

4.2.2 Acoustic contamination

P2 and day 3 of P5 datasets in the reading condition were inspected for acoustic contamination. Philémon Roussel computed correlation matrices between audio and neural signals in 0-200Hz frequency range. To test the statistical significance of the contamination, he then ran a t-test between the mean diagonal correlation of the matrix and a distribution of 10 000 random shuffles of the correlation matrix (Roussel et al., 2020). I decoded mel cepstral coefficients from 21 neural spectral bands for both datasets using a linear regression with 10 frames of both past and future time context. Neural spectral features were selected using a Welch t-test and reduced to 100 features by PCA.

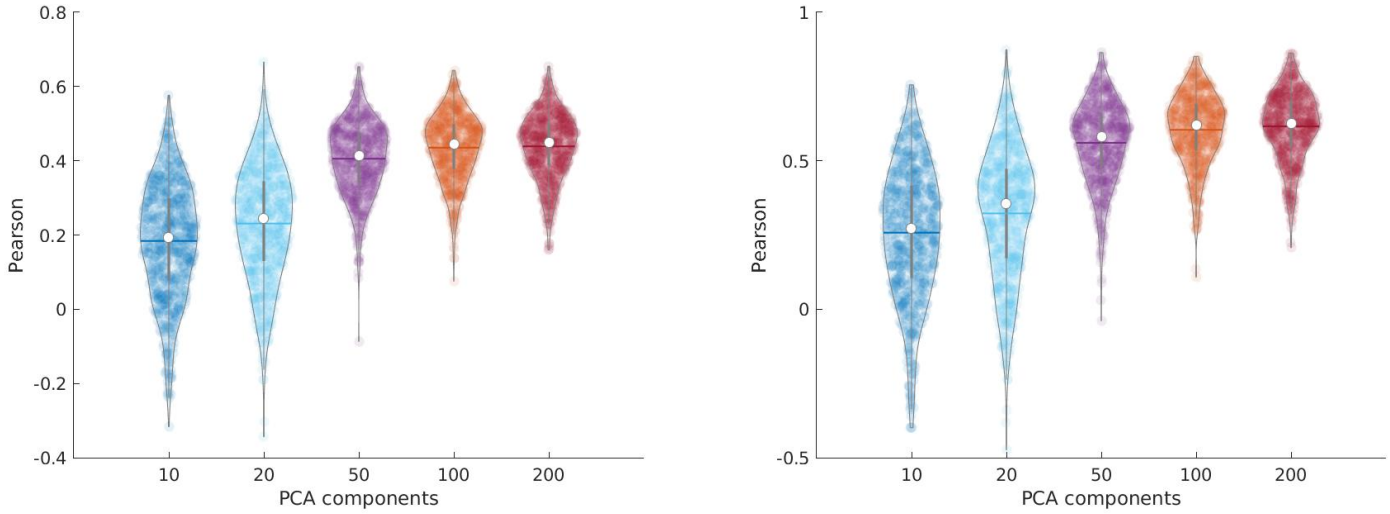
Fig. 3.7 shown in section 3.2.3 already described acoustic contamination in the 0-200Hz range for P2 dataset ($p=0.0$) but not in P5. Some other experiments detailed in the paper showed some contamination in P5 dataset only in a perception set up, never in production. Visual inspection of the correlation matrix shows contamination down to 100Hz in P2. This result coincided with a higher contribution of 90-200Hz spectral range in decoding of P2's mel cepstral coefficients compared to P5, which is shown in Fig. 4.5. In addition to acoustic contamination, decoding of mel cepstral coefficients from P2's neural activity in the 0-90Hz range was poorer than P5's.

4.2.3 Influence of feature reduction on linear decoding

The optimal numbers of reduced features following PCA and PLS were assessed on P5 dataset, including overt speech conditions on day 1,2 and 3. In order to evaluate the quality of the feature reduction, a linear regression with 10 frames of both past and future context (21 total frames, hence 210ms) was computed in a 10-fold crossvalidation. A feature selection was computed over the whole dataset before running the crossvalidation, keeping 1455 features out of 1512.

4.2.3.1. Feature reduction using PCA

In each fold of the crossvalidation, a PCA was computed on the training set's neural features after normalization of the data and before concatenating frames to add temporal context. The first 10, 20, 50, 100 and 200 components were kept to transform both the training and testing sets before training a linear regression to predict F0 and mel cesptral coefficients. The mean mel cesptral coefficients correlation and F0 correlation were computed on each sentence to evaluate the



(a) mel cepstrum

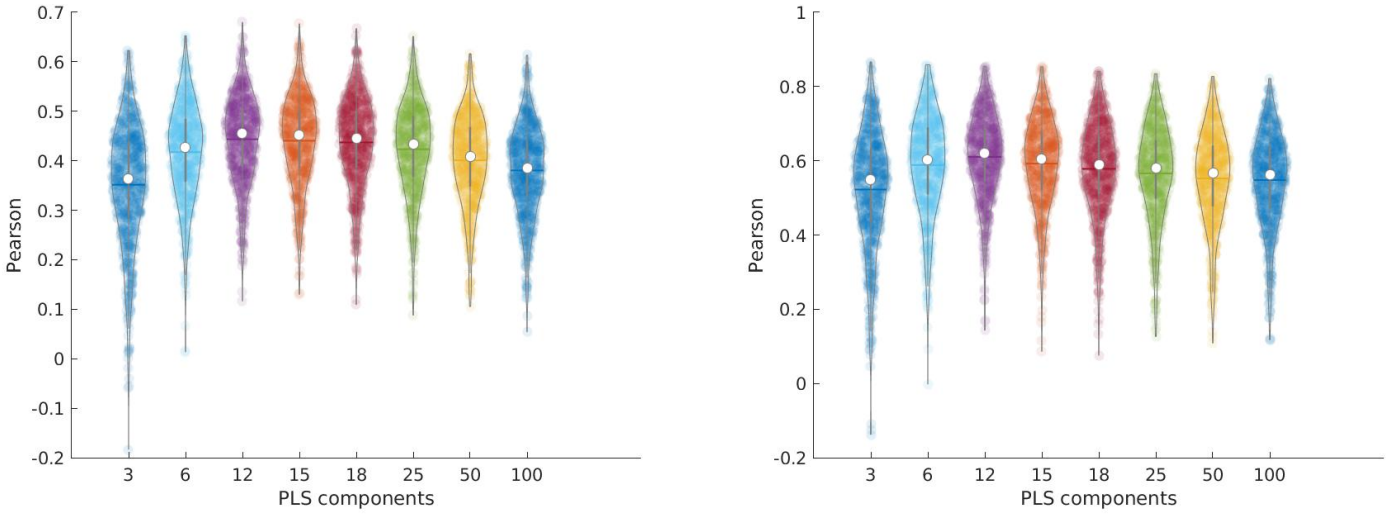
(b) F0

Fig. 4.6.: Linear decoding of speech with varying number of PCA components. Linear decoding of F0 and mel cepstrum of speech after a PCA transformation of neural features with varying number of components. Each point in violin plots represents the average correlation of decoded features on one sentence of the dataset. The white dot shows the median sentence and the colored horizontal line shows the mean correlation.

decoding of speech with the different numbers of PCA components. Fig. 4.6 shows maximum correlations for a 200 PCA components. The correlations increased with the number of PCA components, starting to stagnate around 100 to 200 components.

4.2.3.2. Feature reduction using PLS

In each fold of the crossvalidation, a PLS regression was trained to predict mel cepstral coefficients and F0 from the training set's neural features. Unlike with PCA, the reduction of features computed by the PLS is included in the regression. For this reason, the PLS latent space transformation is computed after both normalization and concatenation of frames to add temporal context. The PLS regression was computed with a latent space of 3, 6, 12, 15, 18, 25, 50 and 100 components. The mean mel cepstral coefficients correlation and F0 correlation were computed on each decoded sentences. The correlations of decoded features for the different number of PLS components are displayed on Fig. 4.7. For both mel cepstral coefficients and F0, the correlations was maximum for 12 components.



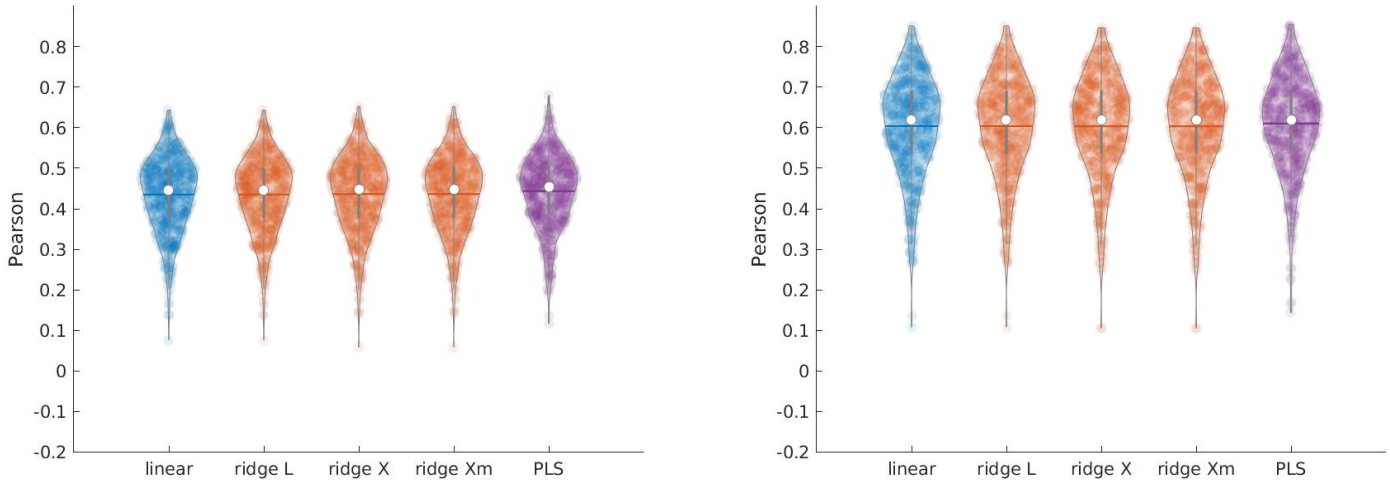
(a) mel cepstrum (b) F0

Fig. 4.7.: Speech decoding by PLS regression with varying number of components on P5 dataset. Decoding of F0 and mel cepstral coefficients of speech from neural features using PLS regression with varying number of components. Each point in violin plots represents the average correlation of decoded features on one sentence of the dataset. The white dot shows the median sentence and the colored horizontal line shows the mean correlation.

4.2.3.3. Comparison of linear methods

Five different linear models were trained to decode F0 and mel cepstral coefficients from neural features on P5 dataset, using 10 frames of both past and future context (21 frames, hence 210ms). A simple linear regression combined with a PCA of neural features keeping the first 100 components was trained to establish a baseline. It was compared to similar linear regression combined with the 3 different ridge regularizations described in section 3.6.2.2: 1. based on L-curve, 2. based on crossvalidation, and 3. based on crossvalidation with individual λ value for each feature (see 3.6.2.2). Finally a PLS regression with 12 components was trained on the same data but without prior PCA on the neural features, unlike the other methods of the comparison. All models were trained in a 10-fold crossvalidation, the correlations of the predicted sentences speech features are displayed in Figure 4.8, showing very similar performance for each method.

All methods performed very closely, with median correlations within 0.02 from eachother on both F0 and mel cepstral coefficients. The best median mel cepstral coefficients correlation was found for the PLS regression at about 0.45, while the F0 correlations for all methods were about 0.62.



(a) mel cepstrum

(b) F0

Fig. 4.8.: Comparison of linear methods for speech decoding on P5 dataset. A linear regression was trained to predict F0 and mel cepstral coefficients from neural features of speech using a PCA with 100 components. Similar ridge regressions were trained using 3 different methods to compute the λ factor: L-curve (L), crossvalidation (X) and crossvalidation with individual λ per features (Xm). Finally, a PLS regression with 12 components was trained to perform the same task.

4.2.4 Additional data

Acoustic features of speech were decoded from EC61 dataset using the PLS regression method designed on P5 dataset. All 9 blocs of EC61 dataset described in section 3.1.2.4 were cleaned by common median reference before extracting spectral features from neural signals (see section 3.2.2.1). As EC61 sentences are not part of BY2014, it was not possible to estimate articulatory trajectories from audio recordings. Only direct decoding paradigm was therefore applied to EC61 dataset in this section.

PLS regressions were trained to predict mel cepstral coefficients and F0 from 21 spectrogram features of neural activity. Neural features were concatenated to form 10 frames of both past and future context after selecting 1500 out of 5376 (256 electrodes*21 frequency bands) features with a Welch *t*-test (see figure 3.2.4). The feature selection was necessary to fit matrices in memory, some preliminary experiments showed that reducing the number of selected features hindered decoding.

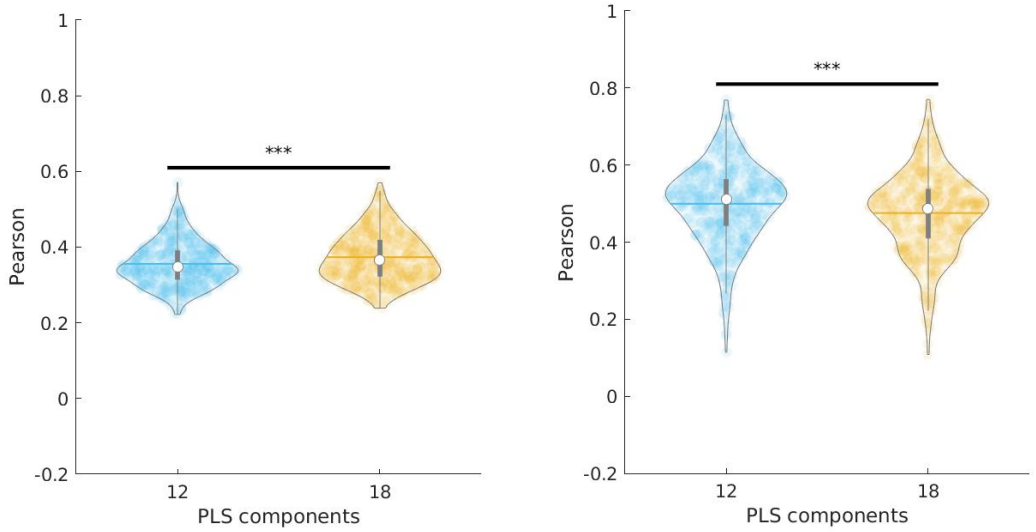
Correlations of decoded mel cepstral coefficients and F0 with ground truth are shown on figure 4.9, for PLS regressions with 12 and 18 components. Best mel cepstral coefficients correlations were found for 18 components with a median correlation of 0.36, and best F0 correlations were found for 12 PLS components with a median correlation of 0.51.

4.2.5 Influence of context and delays

Because PLS gave best performance and did not require any initial step of feature reduction, we used PLS models with 12 components to test the influence of context size and delay on speech decoding from neural activity in 10-fold crossvalidations.

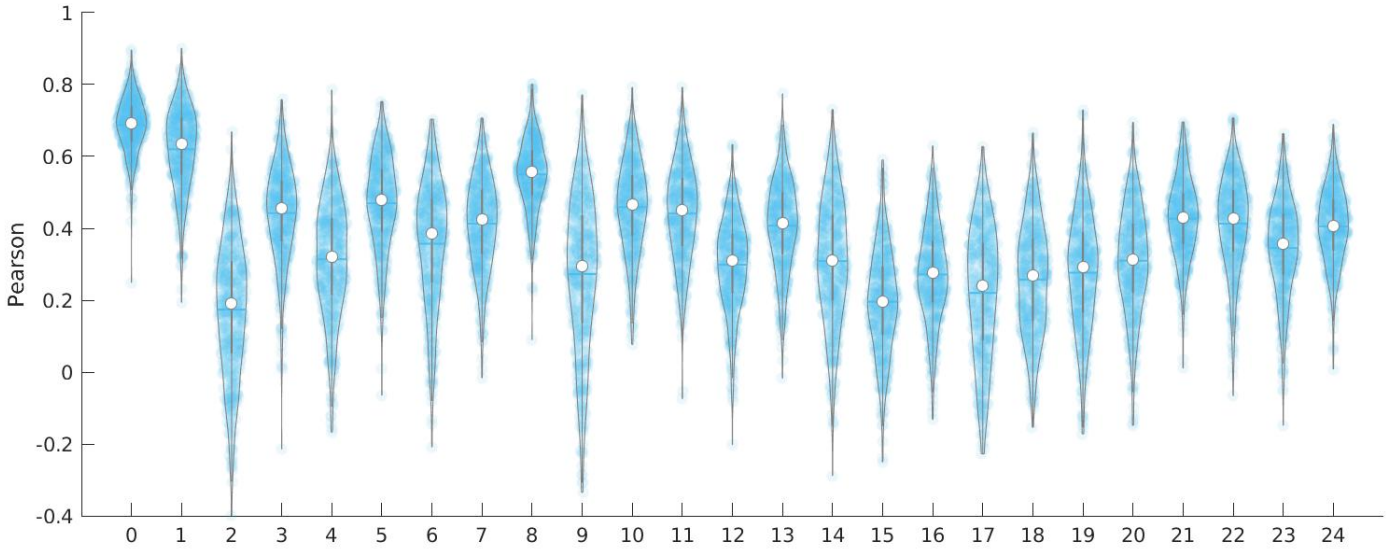
4.2.5.1. Context size

Before computing PLS regressions, various temporal contexts were added to neural features so that multiple frames of neural features were used to decode a single speech frame. The decoded F0 and mel cepstral coefficients correlations are displayed in Fig. 4.10 for context size of 0, 5, 11 and 21 frames (1 frame = 10ms). A context size of 0 means that one frame of neural features was used to decode the corresponding frame of speech features, while a context size of 5 frames means that 2 past frames and 2 future frames were concatenated to decode the same frame of speech features.



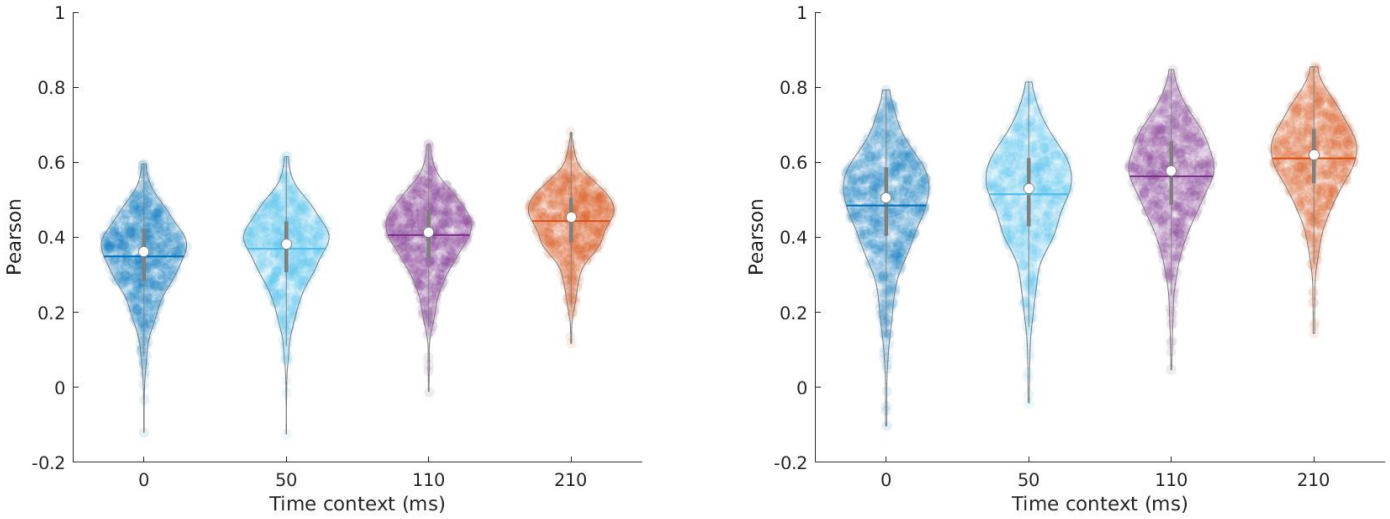
(a) Mel cepstrum

(b) F0



(c) Mel cesptral coefficients (12 PLS components)

Fig. 4.9.: Decoding of acoustic features of speech with PLS regression from EC61 dataset. PLS regressions with 12 and 18 components were trained to predict mel cepstral coefficients and F0 from neural features. Statistical significance of decoding with 12 vs 18 components was assessed by a Wilcoxon signed-rank test. (a) Correlations of decoded mel cepstrum with matching ground truth using 12 and 18 PLS components (median correlations $r_{12} = 0.35$, $r_{18} = 0.36$, $p = 3.3e^{-35}$). (b) Correlations of decoded F0 with matching ground truth using 12 and 18 PLS components (median correlations $r_{12} = 0.51$, $r_{18} = 0.49$, $p = 3.0e^{-29}$). (c) Correlations of individual mel cepstral coefficients predicted with the 12 components PLS.
n.s.: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$



(a) mel cepstrum

(b) F0

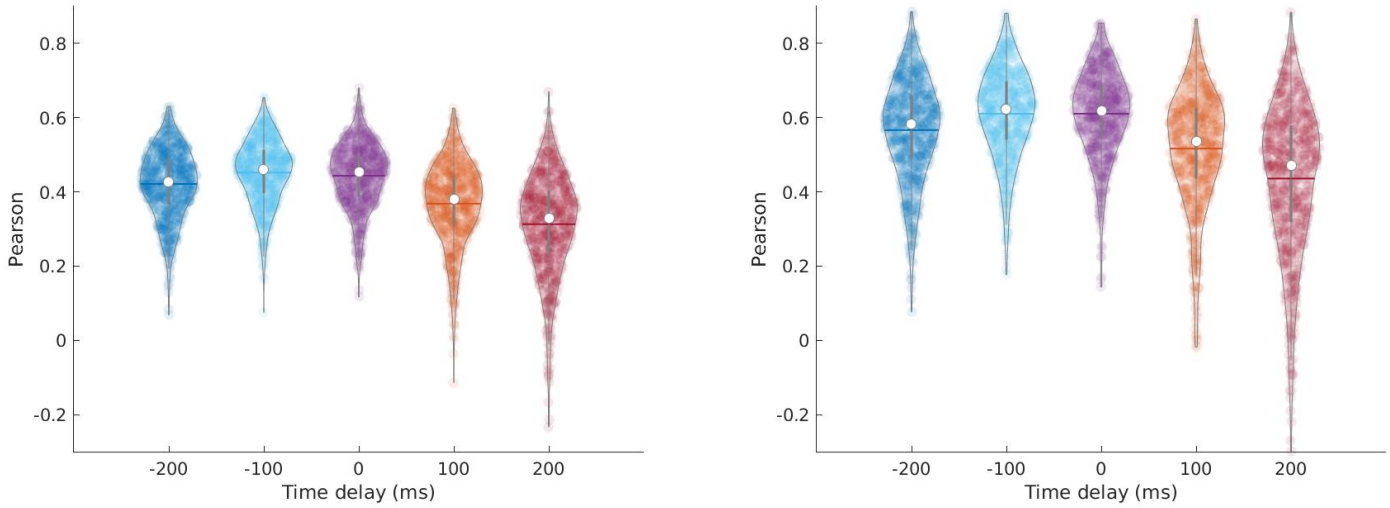
Fig. 4.10.: PLS decoding of speech from neural activity with varying temporal contexts on P5 dataset. A PLS regression predicted F0 and mel cepstral coefficients from neural features with 0 time delay and varying temporal contexts (*ie.* time interval around time of prediction). On the horizontal axis, the total number of frames used as context, centered around the current frame: half before and half after.

Decoding correlations improved with the context size, reaching a maximum for a context size of 21 frames (hence 210ms) with median correlations of 0.45 for average mel features and 0.62 for F0.

4.2.5.2. Time delays

Using the larger temporal context of 21 frames (210ms), PLS models were trained to decode speech from neural features after applying a time delay to neural features compared to speech features. Decoding with time delays $d = -200, -100, 0, 100, 200$ (in ms) were investigated, the resulting correlations are shown in Fig. 4.11. For a time delay d, the PLS model predicts a frame of speech features at time t from neural features at time $t + d - 100, \cdot, t + d, \cdot, t + d + 100$.

For all speech features, decoding reached a maximum for a time delay of $d = -100ms$ (10 frames), *i.e.* using frames from time $t - 200ms$ to t to decode a frame of speech features at time t, with median decoding correlations of 0.46 for mel cepstrum and 0.62 for F0. Although using no time delay at all reached a very

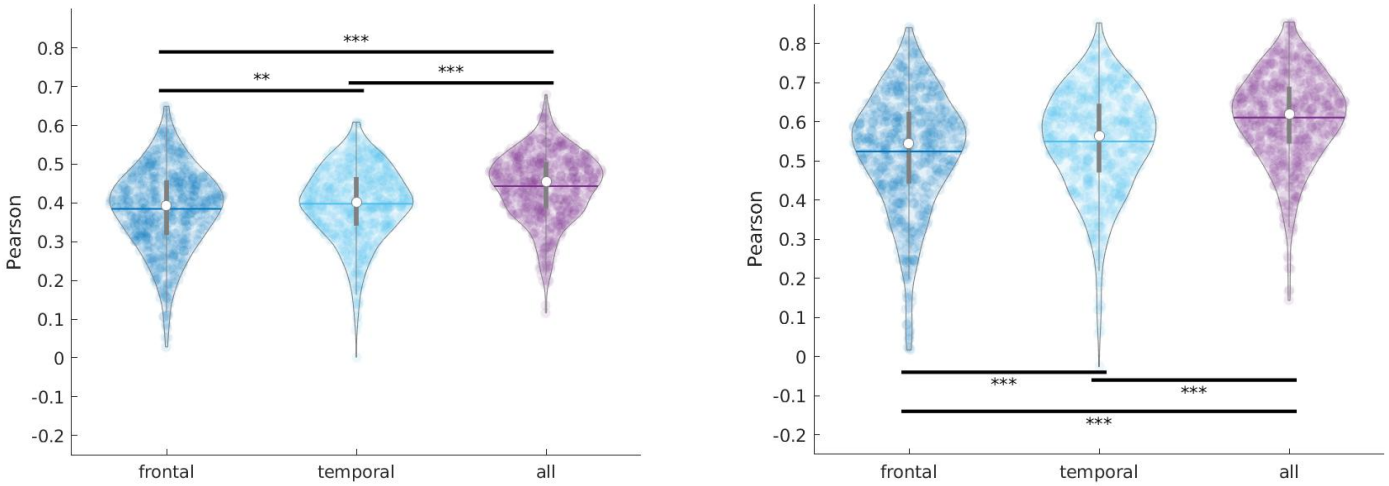


(a) mel cepstrum

(b) F0

Fig. 4.11.: PLS decoding of speech from neural activity with varying time delays on P5 dataset. A PLS regression predicted F0 and mel cepstral coefficients from neural features with 10 frames of both past and future context. Different regressions were computed by changing the time alignment between neural and decoded features. A negative delay means that decoded features are predicted from past neural activity, while positive delays implies that decoded features are predicted from future neural activity.

close performance, adding more positive delay decreased more and more decoding performance.



(a) Mel cepstrum: direct decoding

(b) F0

Fig. 4.12.: Comparison of frontal and temporal electrodes for decoding of speech from P5 dataset. (a) Pearson correlations of mel cepstral coefficients decoded by a PLS regression with 12 components from either frontal, temporal or all electrodes (median correlations r=0.39, r=0.40, r=0.45 for frontal, temporal and all electrodes respectively). (b) Decoding of F0 (r=0.54, r=0.56, r=0.62). Wilcoxon signed rank test significance - ***: p<0.001, **: p<0.01

4.2.6 Influence of frontal and temporal activity

In order to test the influence of critical areas of speech decoding, acoustic features of speech were directly decoded from P5 neural features from only frontal or temporal electrodes according to the delimitation shown in section 3.2.2.2. A PLS regression with 12 components was trained in each case and compared to the case where all electrodes were considered.

Pearson correlations of decoded with ground truth features are displayed in figure 4.12 for P5 dataset, statistical significance was computed using Wilcoxon signed-rank tests. First, decoding with all electrodes increased correlations compared to isolated frontal or temporal electrodes for both mel cepstrum ($p = 2.0e^{-61}$ and $p = 1.0e^{-61}$, respectively) and F0 ($p = 8.6e^{-63}$ and $p = 1.4e^{-59}$, respectively). Second, direct decoding of mel cepstral coefficients and F0 performed slightly better with temporal electrodes than with frontal electrodes ($p = 0.0028$ and $p = 5.8e^{-5}$, respectively).

4.3 Articulatory Synthesis

In order to test indirect speech decoding, an articulatory-to-acoustic speech synthesizer is required. Previous work by Bocquelet et al. (2016c) used a non optimized feedforward DNN. We investigated here other architectures to target improved synthesis. Regression models based on recurrent and feedforward artificial neural network architectures were trained to predict mel cepstrum of speech from articulatory trajectories. The quality of the articulatory-to-acoustic synthesis was assessed for each network using both an objective criterion and subjective listening tests.

All the work presented in this section has been mostly done under the direction of Thomas Hueber and Laurent Girin in Gipsa-Lab, Grenoble.

4.3.1 Objective comparison

4.3.1.1. Comparison of temporal contexts

Deep neural networks (DNN) with varying past and future temporal contexts $(\tau_p, \tau_f) \in [0, 1, 3, 5, 7, 9]^2$ (in frames; 1 frame = 5ms) and a bidirectional Gated Recurrent Unit (biGRU) were trained to perform articulatory-to-acoustic regression on 3 electromagnetic articulography datasets: BY2014, mocha and PB2007 (see section 3.1.1). While DNNs use finite temporal contexts, the biGRU should by design use indefinitely long temporal context (in practice complete sentences).

Compared to the other experiments focused on speech decoding from cortical activity, the following experiments used a slightly different processing of speech: the mel-cepstral analysis was performed at a 200Hz rate instead of 100Hz, with 26 mel-cepstral coefficients instead of 25. Additionally, as recordings usually leave some silence at every beginning and end of sentences. To make sure silences were not overly represented in the dataset compared to speech, silences were trimmed to be only 50ms long before and after speech. Finally, out of 676 sentences, only the first 533 sentences of BY2014 were used due to an unexplained drop of performance on the last sentences. All dataset's articulatory features were sampled at 200Hz to match mel-cepstral coefficients, using linear interpolation if necessary. The data is summarized in the following table:

	BY2014	Mocha	PB2007
language	french	southern english	french
speaker	male	male	male
features	27	14	12
sentences	532	460	1109

Neural networks were trained and evaluated in a 5-fold crossvalidation where 3 parts were used for training, 1 for validation and early stopping, and 1 for evaluation. Each model was implemented with Keras (Chollet et al., 2015) and trained on a GPU cluster using Adam optimizer (Kingma and Ba, 2014), mean squared error as loss function, tanh activations, 25% dropout (Srivastava et al., 2014), a batch size of 32, and in the case of DNNs batch normalization (Ioffe and Szegedy, 2015). Training was automatically stopped by early stopping with a patience of 20 epochs to ensure no overfitting. After a grid search of hyperparameters, DNNs were built with 3 hidden layers of 512 neurons, and the biGRU was built with a single layer of 1024 neurons.

Performance of each model was reported as mean squared error shown in Figure 4.13. In all three datasets, the best performance was achieved by a DNN using the maximum past temporal context (45ms), beating the biGRU. Although the number of future temporal context frames necessary to achieve best performance changes from one dataset to another (BY2014: 5ms, P2007: 45ms, MOCHA: 15ms), at least some future context was always used.

4.3.1.2. Real-time compatibility

From the previous results described in section 4.3.1.1, the DNNs evaluated on BY2014 with 10 frames of past context and varying future context sizes were compared using mel cepstral distortion (defined in section 3.5.3). Like on other datasets, adding future temporal context improves articulatory-to-speech synthesis on BY2014, although using only two frames already reached the best performance. The plots in Figure 4.14 show that by adding more future context, performance mostly stagnates, when adding only one frame does give a significant improvement compared to no future context at all.

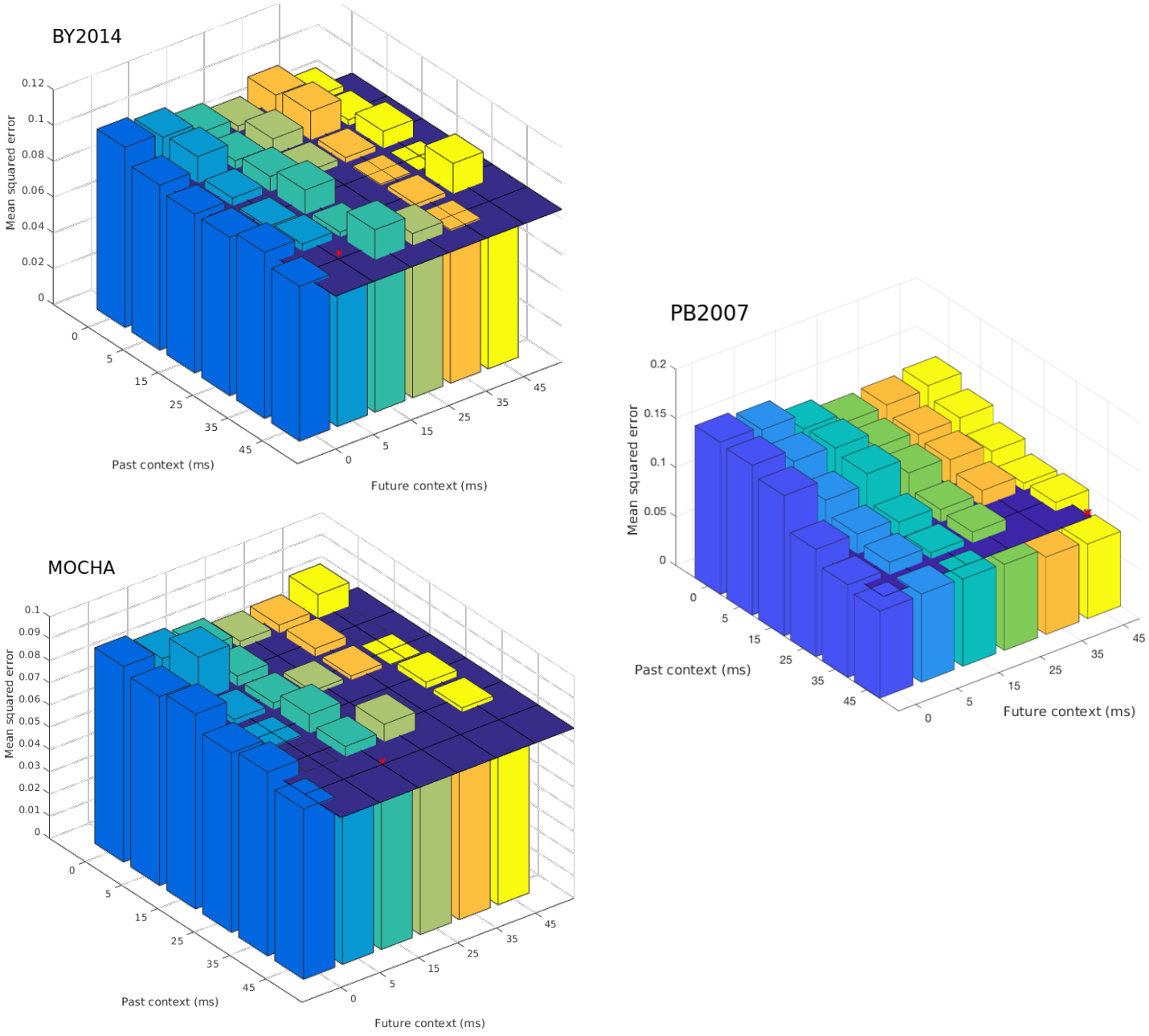


Fig. 4.13.: Mean squared error of articulatory-to-acoustic synthesis of BY2014 (top left), mocha (bottom left) and PB2007 (right) - *the less the better.* DNNs (3 hidden layers of 512 units) MSE scores are displayed by bar plots for each time context, whereas the blue plan indicate the MSE score of the GRU (1 hidden layer of 1024 units). Best DNNs are signaled by a red star.

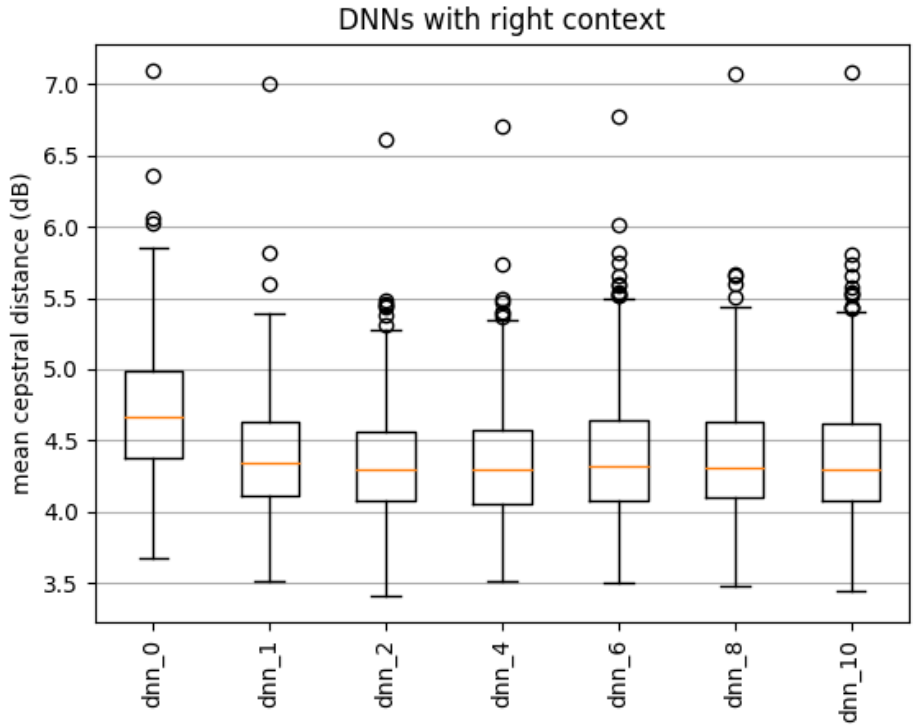
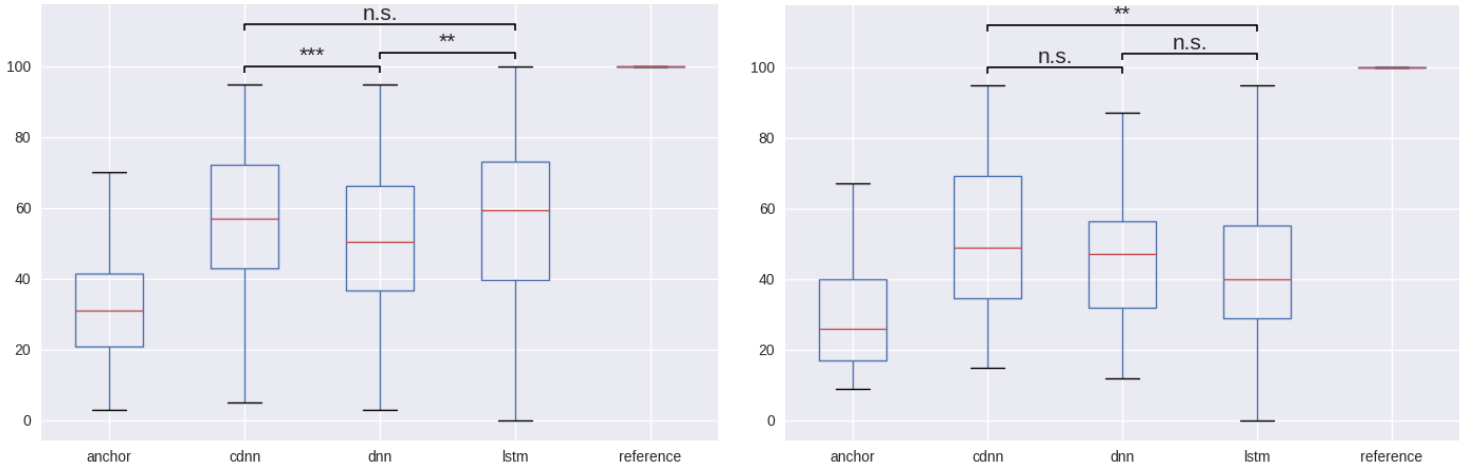


Fig. 4.14.: Influence of future context for articulatory-to-acoustic synthesis with DNNs. 7 DNN models are compared, all trained on BY2014 to predict mel cepstrum from articulatory trajectories. From left to right, box plots show the mel cepstral distortion over all sentences of BY2014 for DNNs with 0 to 10 frames of future context. All models also used 10 frames of past context.

4.3.2 Subjective evaluation

A MUSHRA test described in section 3.5.4 was setup to evaluate the general quality of articulatory-to-acoustic speech synthesis from 3 neural networks already described in section 3.5.2: a biGRU, a DNN without temporal context, and a DNN with 10 frames of past context and 1 frame of future context. The test was made freely accessible online from a generic link which was shared among other students. An initial pilot study was used to design a low anchor that provides good enough synthesis so that other methods ratings are not overly squashed, finally settling down on a feedforward DNN with one hidden layer of 32 neurons. Participants were asked to rate 9 vowel-consonant-vowel (VCV) sequences and 10 randomly selected sentences from BY2014.

Six self assessed native french speakers (3 males, 2 females, 1 other) aged from 20 to 31 completed the test, none of whom had to be excluded. The rating scores of each models, anchor and reference were computed for each stimulus and participant



(a) Sentences

(b) VCV

Fig. 4.15.: Subjective evaluation of synthesized sentences from BY2014. (a) MUSHRA evaluations of complete sentences synthesized by DNN, DNN with context, LSTM as well as reference and low anchor (b) MUSHRA evaluations of VCVs synthesized by DNN, DNN with context, LSTM as well as reference and low anchor (example VCV: *Tu t'appelles apa c'est ça?*)
Wilcoxon signed rank test significance - ***: $p<0.001$, **: $p<0.01$, n.s.: $p>0.05$.

and displayed as box plots in Figure 4.15. Difference of performances between the different models were assessed by Wilcoxon signed-rank test. All models were clearly rated higher than the low anchor and well below the reference, in the 20 to 80 bracket as expected for a correctly designed MUSHRA test. Both the biGRU and the DNN with temporal context performed seemingly higher than the DNN without temporal context on sentences ($p=0.0012$ and $p=0.00018$, respectively), while no significant difference of performance was found between the biGRU and the DNN with context ($p=0.68$). On VCVs alone, the biGRU performed worse than the DNN with context ($p=0.0077$) but no significant difference was found between the DNN with context and the DNN without context ($p=0.089$) or between the DNN without context and the biGRU ($p=0.19$).

4.4 Indirect decoding of speech through an articulatory representation

Acoustic features of speech were previously decoded from neural activity using linear models. This section investigates the decoding of articulatory trajectories and the

indirect decoding of mel cepstral coefficients through these decoded articulatory trajectories, as detailed in section 3.6.3.2.

Articulatory-to-acoustic regressions based on Recurrent Neural Networks did not prove to be more efficient than simple feedforward neural networks. However adding past context and some future context improved decoding on all datasets. Therefore the model that was used in this section is a feedforward neural network with 3 hidden layers of 512 neurons, 10 frames of past context and 1 frame of future context. This introduces a small latency of 10ms, which should not cause too many issues for a real-time use (Lee, 1950; Stuart et al., 2002) while improving the overall decoding as shown in section 4.3.1.2. Training was done using 25% dropout, batches of size 32, Adam optimizer, and early stopping with a patience of 20 epochs.

In all of the following sections, P5 articulatory trajectories were inferred from BY2014 using dynamic time warping for days 1,2 and 3 on overt conditions (see 3.4.2). Those articulatory trajectories were then decoded from neural activity by linear methods in a 10-fold crossvalidation (see section 3.6.5.1).

4.4.1 Decoding of articulatory trajectories

The decoding of articulatory trajectories from neural activity was investigated using linear methods on days 1,2,3 of P5 dataset.

4.4.1.1. Comparison of linear methods for articulatory trajectories decoding

Five different linear models were trained to decode articulatory trajectories from neural features on P5 dataset, using 10 frames of both past and future context. A simple linear regression combined with a PCA of neural features keeping the first 100 components was trained to establish a baseline. It was compared to similar linear regression combined with the 3 different ridge regularizations described in section 3.6.2.2: 1. based on L-curve, 2. based on crossvalidation, and 3. based on crossvalidation with individual λ value for each feature. Finally a PLS regression with 12 components was trained on the same data but without prior PCA on the neural features, unlike the other methods of the comparison. All models were trained in a 10-fold crossvalidation, the correlations of the predicted sentences speech features are displayed in Figure 4.16. All methods performed very closely, with median correlations within 0.01 from eachother. The best median articulatory

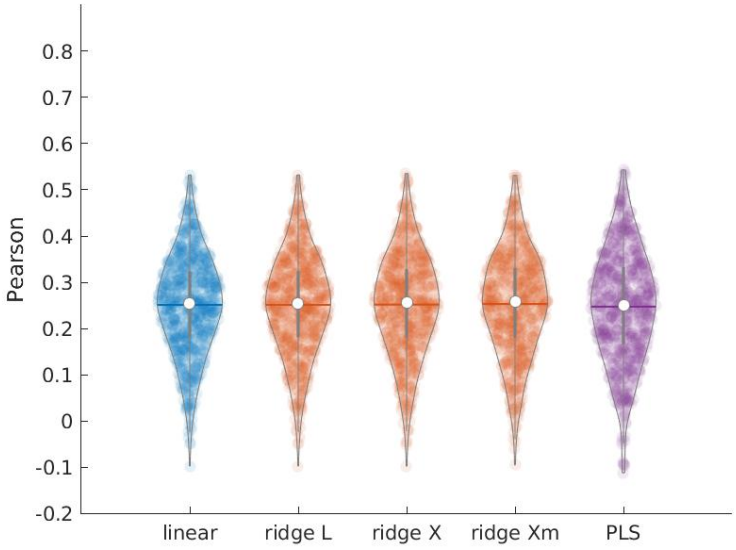


Fig. 4.16.: Comparison of linear methods for decoding of articulatory trajectories on P5 dataset. A linear regression was trained to predict articulatory trajectories from neural features of speech using a PCA with 100 components. Similar ridge regressions were trained using 3 different methods to compute the λ factor: L-curve (L), crossvalidation (X) and crossvalidation with individual λ per features (Xm). Finally, a PLS regression with 12 components was trained to perform the same task.

features decoding was performed by ridge regressions using crossvalidation with a median correlation of 0.26. However, because we did not find any statistically significant differences across these linear methods, we further used PLS, which was the chosen method for direct decoding.

4.4.1.2. PLS decoding of articulatory trajectories

A PLS regression was trained to predict P5's articulatory trajectories from neural features. Neural features were then concatenated with 10 frames (100ms) of both past and future context before running the PLS model with 12 components, hence 200ms of neural data. The Pearson correlations of individual articulatory trajectories were computed over each sentence of the dataset. Chance levels were estimated by randomly shuffling the samples of neural data and running the complete decoding pipeline in the exact same way. Both chance and decoding correlations are shown on Figure 4.17 for each articulator in both caudo-rostral (horizontal front to back, x) and ventro-dorsal (vertical down to up, y) axes.

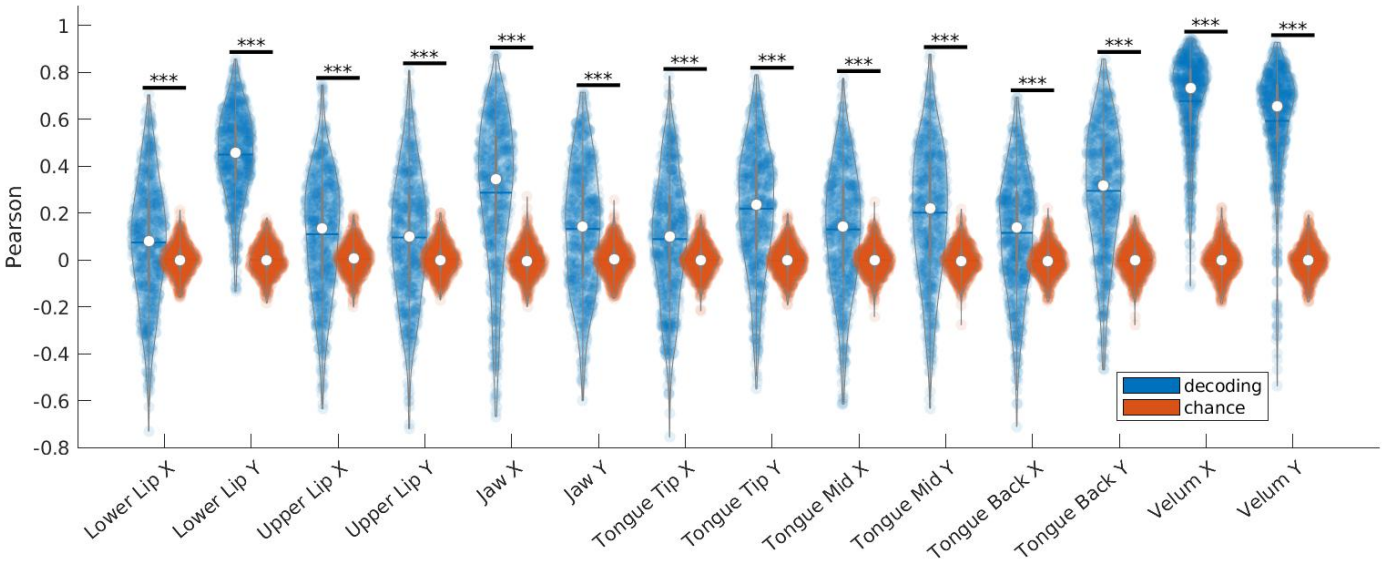


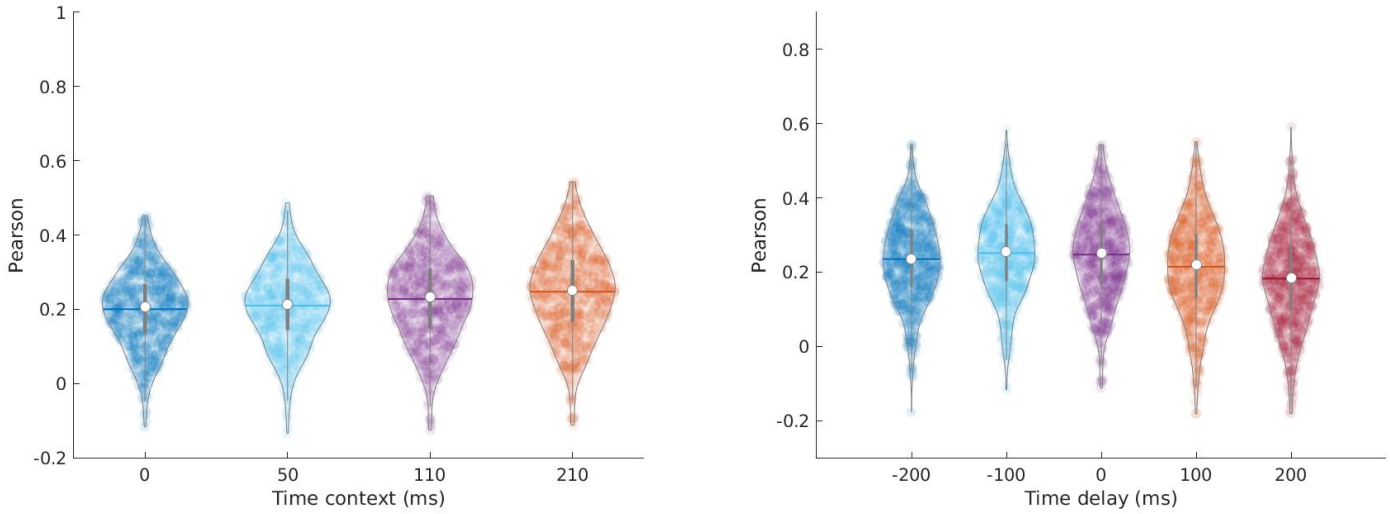
Fig. 4.17.: Linear decoding of P5 estimated articulatory trajectory. Correlations of individual decoded articulatory trajectories (in blue) for both horizontal (X) and vertical (Y) axis and their corresponding chance levels (in red).
***: Wilcoxon signed-rank test p-value < 0.001

All decoded articulatory trajectories were largely different from chance levels (Wilcoxon signed rank test, p-values all below $1e10^{-10}$), however the amplitude of correlation varied vastly from one sentence to another for most articulators. The decoding correlations of the *velum* was high for both axes, reaching a median of 0.73 for the caudo-rostral axis. The second best decoding was achieved for the vertical axis of the lower lip, followed by the horizontal axis of the jaw. The vertical axis of the three tongue points were also decoded to some extent with median correlations from 0.22 (mid) to 0.32 (back). Other articulatory trajectories were poorly decoded with median correlations around 0.1, including the upper lip, horizontal axis of the tongue, vertical axis of the jaw and horizontal axis of the lower lip.

4.4.1.3. Influence of time context and delays on decoding of articulatory trajectories

PLS models with 12 components were trained to predict articulatory trajectories from neural features on P5 dataset using several different time contexts and time delays. The different regressions were evaluated in 10-fold crossvalidations.

Before computing PLS regressions, various temporal contexts were added to neural features so that multiple frames of neural features were used to decode a single



(a) Influence of time context

(b) Influence of time delay

Fig. 4.18.: PLS decoding of articulatory trajectories from neural activity with varying temporal contexts on P5 dataset. A PLS regression predicted articulatory trajectories from neural features with varying temporal contexts (*ie.* time interval around time of prediction). On the horizontal axis, the total number of frames used as context, centered around the current frame.

speech frame. The decoded articulatory trajectories correlations are displayed in Fig. 4.18 for context size of 0, 5, 11 and 21 frames (1 frame = 10ms). A context size of 0 means that one frame of neural features was used to decode the corresponding frame of speech features, while a context size of 5 frames means that 2 past frames and 2 future frames were concatenated to decode the same frame of speech features. Decoding correlations improved with the context size, reaching a maximum for a context size of 21 frames (hence 210ms) with median correlations of 0.25 for average articulatory features.

Using the larger temporal context of 21 frames (210ms), PLS models were trained to decode articulatory trajectories from neural features after applying a time delay to neural features compared to articulatory features. Decoding with time delays (in ms) $d = -200, -100, 0, 100, 200$ were investigated, the resulting correlations are shown in Fig. 4.18. For a time delay d, the PLS model predicts a frame of articulatory features at time t from neural features at time $t + d - 100, \cdot, t + d, \cdot, t + d + 100$.

Decoding reached a maximum for a time delay of $d = -100ms$ (10 frames), *i.e.* using frames from time $t-200ms$ to t to decode a frame of articulatory trajectories at time t, with median decoding correlations of 0.25 for articulatory features. Although

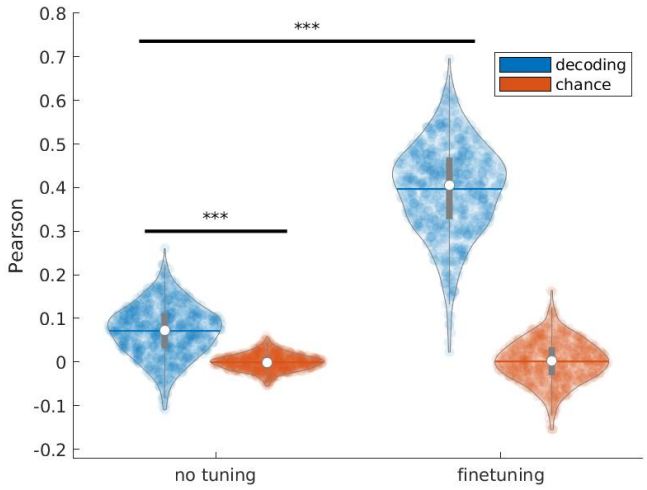


Fig. 4.19.: Effect of finetuning of articulatory synthesis for indirect decoding (P5 dataset). Correlations of indirectly decoded mel cepstral coefficients (blue) are compared to their matching chance level (red). Left violin plots report evaluation of indirect decoding without finetuning, and right violin plots report evaluation of the same decoder finetuned on P5 data.
***: Wilcoxon signed-rank test p-value < 0.001

using no time delay at all reached a very close performance, adding more positive delay decreased more and more decoding performance.

4.4.2 Transfer learning of DNN-based articulatory-to-acoustic synthesis

This section investigates the prediction of mel cepstral coefficients from decoded articulatory trajectories on P5 dataset using a pretrained articulatory-to-acoustic DNN. The resulting indirect decoding of mel cepstral coefficients was compared with and without finetuning on P5 dataset.

The articulatory-to-acoustic neural network was trained on BY2014 using early stopping and a crossvalidation according to the methodology of section 3.5.3.1. The best model was saved and used to predict mel cepstral coefficients from decoded articulatory trajectories of P5, which were decoded from neural activity using a PLS regression with 12 components (see 4.4.1.2). It was also **finetuned** on P5's dataset according to section 3.6.3.2 to improve the prediction of P5's mel cepstral coefficients from decoded articulatory trajectories by transfer learning.

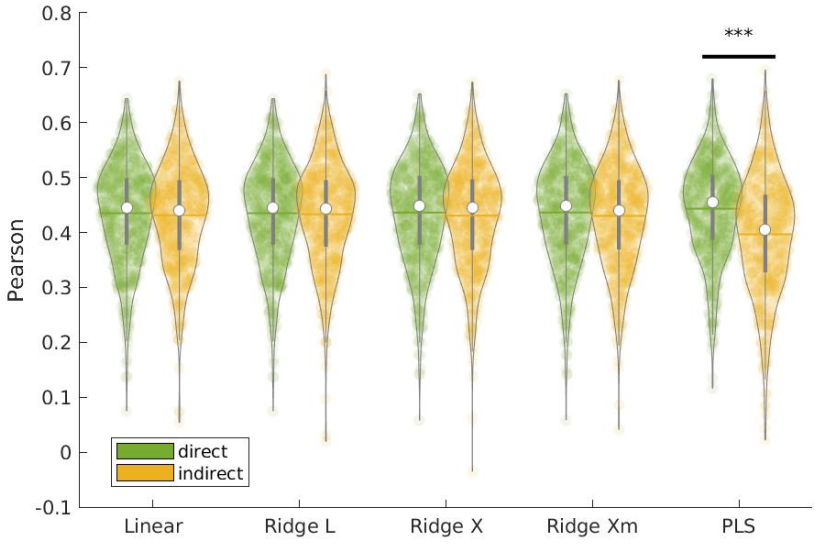
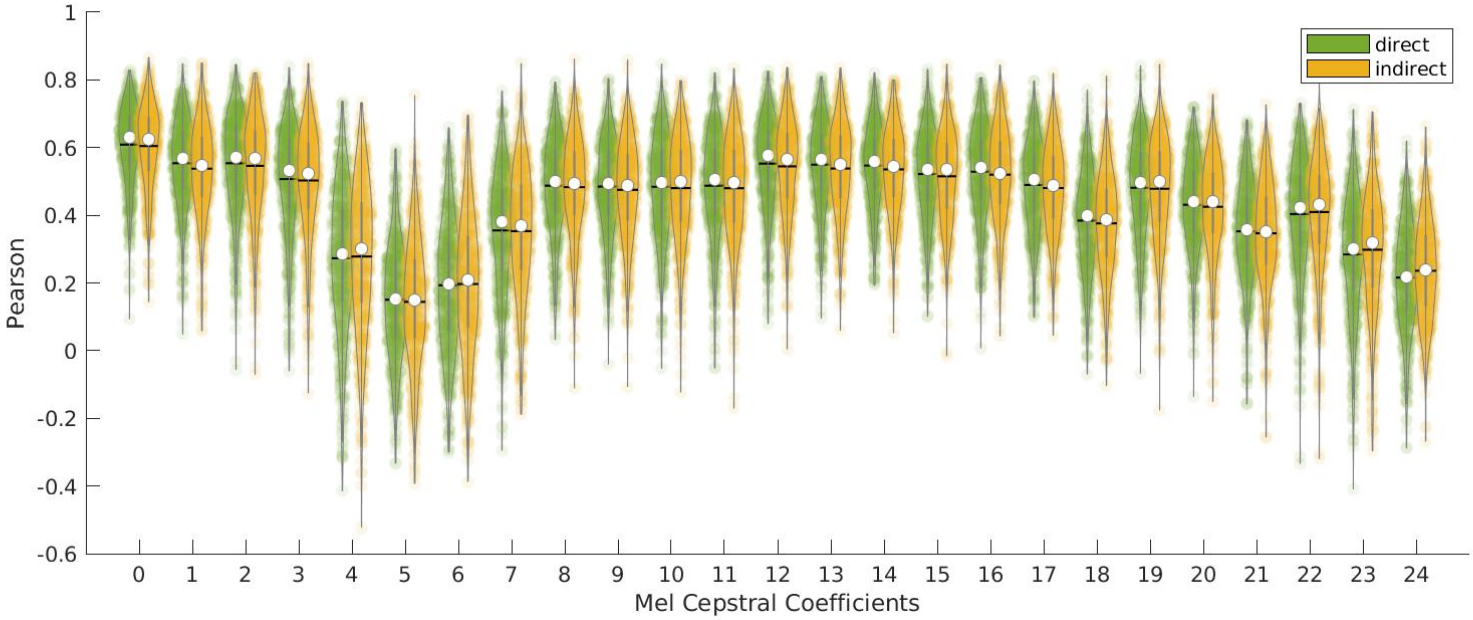


Fig. 4.20.: Comparison of direct and indirect decoding of mel cepstral coefficients on P5 dataset. Left violin plots (green) show correlations of mel cesptral coefficients directly decoded from neural features using linear, ridge and PLS regression. Right violin plots (yellow) show correlations of mel cepstral coefficients predicted by an articulatory-to-acoustic DNN from articulatory trajectories decoded by the same linear methods. Linear and ridge regression were trained on 100 PCA components and the PLS regression used 12 components. Ridge regressions were trained using 3 different methods to compute the λ factor: L-curve (L), crossvalidation (X) and crossvalidation with individual λ per features (Xm). ***: Wilcoxon signed-rank test p-value < 0.001

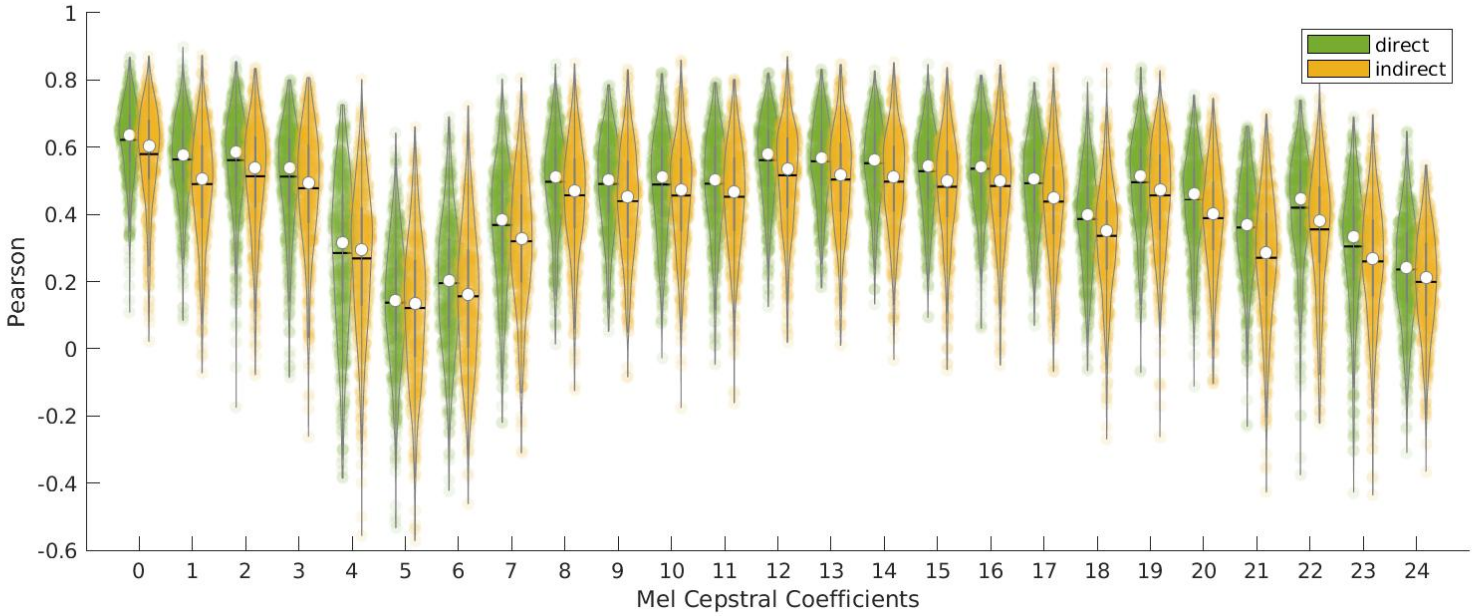
Correlations of predicted mel cepstral coefficients with or without finetuning were computed with their matching ground truth as described in section 3.6.5.2. A chance level was computed by shuffling all samples of neural data and running the same decoding pipelines. Results are reported in figure 4.19, showing a large improvement of decoding when using finetuning (finetuning: 0.41, no tuning: 0.07; median reported correlations). Both methods achieved better performance than chance, even without finetuning.

4.4.3 Comparison of direct vs indirect decoding of acoustic speech features

The direct decoding of mel cepstral coefficients using linear methods was compared with the indirect decoding of mel cepstral coefficients through and articulatory representation (see 3.6.3). Two linear regression methods were considered for the decoding of articulatory trajectories and direct decoding of mel cepstral coefficients:

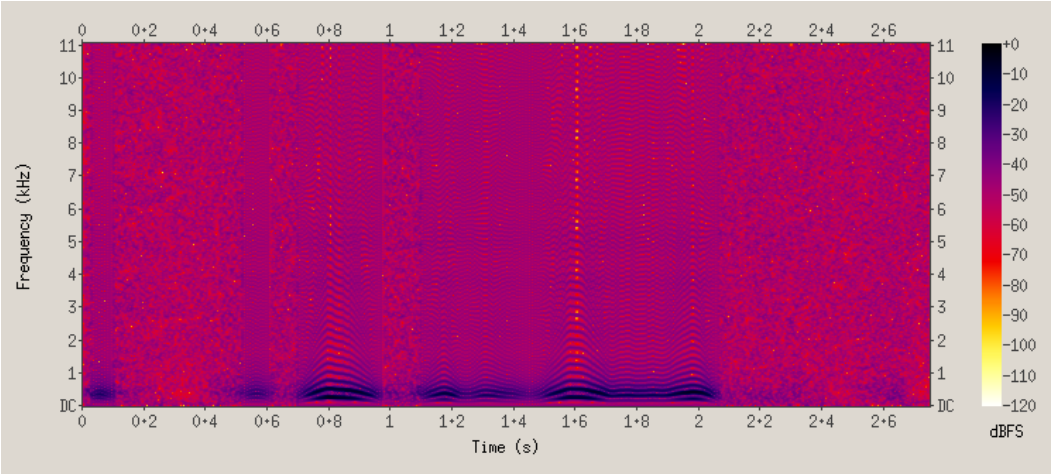


(a) Linear

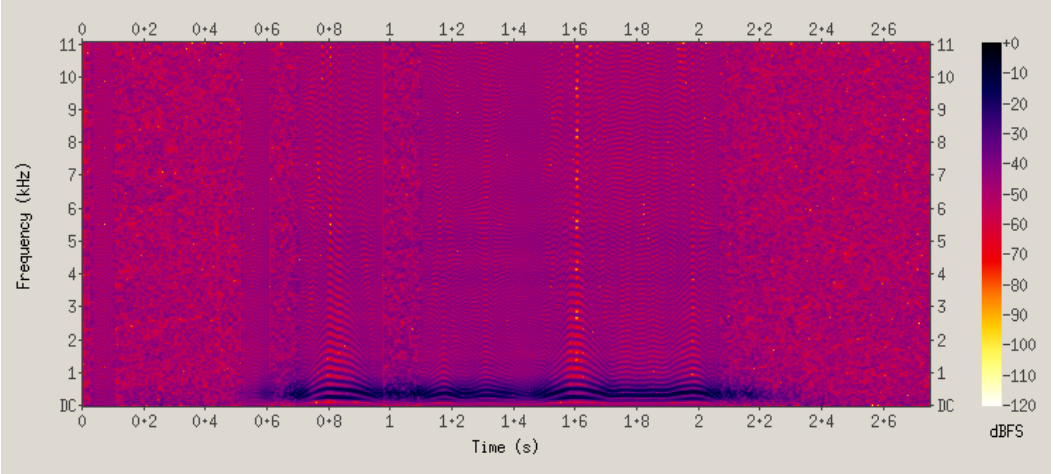


(b) PLS

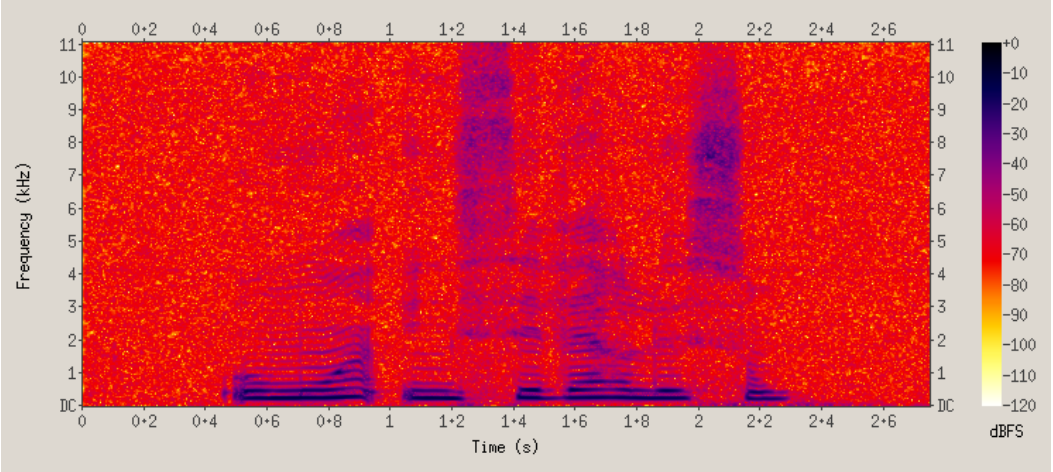
Fig. 4.21.: Comparison of mel cepstral coefficients decoded through direct and articulatory paradigms for P5 dataset. (a) correlations of decoded and extracted mel cepstral coefficients using linear regression with 100 PCA components. Each coefficient's left violin plot shows the direct prediction correlations while the right violin plot shows the articulatory-based prediction correlations. Median correlation is represented by a white dot and mean correlation is represented by an horizontal line. (b) same as *a* but using a PLS regressions with 12 components.



(a) Direct



(b) Indirect



(c) Original

Fig. 4.22.: Spectrograms of decoded sentence from P5 dataset: *'Au moins une sévère leçon.'* (a) Directly decoded from cortical activity using PLS regression with 12 components. (b) Indirectly decoded from cortical activity using PLS regression with 12 components and the articulatory-to-acoustic DNN. (c) Original sentence recorded in P5 dataset.

1. a linear regression combined with a PCA reduction of neural features by keeping the first 100 components and 2. a PLS regression with 12 components. Both methods used 210ms of temporal context and 0ms of time delay. The articulatory-to-acoustic synthesizer described in the preamble of section 4.4 was used to predict mel cepstral coefficients from decoded articulatory trajectories after finetuning (as in 4.4.2).

Mean correlations between decoded and actual mel cepstral coefficients were computed across all coefficients and are shown in figure 4.20. Indirect decoding of speech using linear and ridge decoders performed close to direct decoding, while never achieving higher mean or median correlations. However, indirect decoding using PLS regression performed significantly lower than direct decoding ($p = 5e10^{-37}$, Wilcoxon signed-rank test) with respective median correlations of 0.41 and 0.45.

Correlations between each decoded and actual mel cepstral coefficients were compared for both direct and indirect predictions in figure 4.21. They were similar for direct and indirect decoding using linear regression, while indirect correlations were lower than direct correlations using PLS regression. Example spectrograms of a decoded sentence using both direct and indirect decoding with the PLS regression is shown in Fig. 4.22. Decoded sentences were synthesized with an MLSA filter (see 3.6.4.1). On this example, both methods show a similar performance, with a poor decoding of higher frequencies. Both decoding models seem to predict average values of formants but do not track their trajectories.

4.4.4 Influence of Neural features

4.4.4.1. High gammas

Acoustic and articulatory features of speech were decoded from P5 neural features with and without high gamma bands above 90 Hz. Decoding was performed by a linear model with 100 PCA components for direct prediction of acoustic features and an articulatory-to-acoustic DNN for indirect decoding of mel cepstral coefficients.

The decoding results are presented in figure 4.23. While speech decoding methods often relies on high gamma frequencies of neural activity, correlations between decoded and ground truth speech features showed a relatively small decrease when removing high gamma frequencies above 90 Hz from decoding features. Adding high gamma features nonetheless significantly increased decoding for all speech features with both the direct and indirect paradigms.

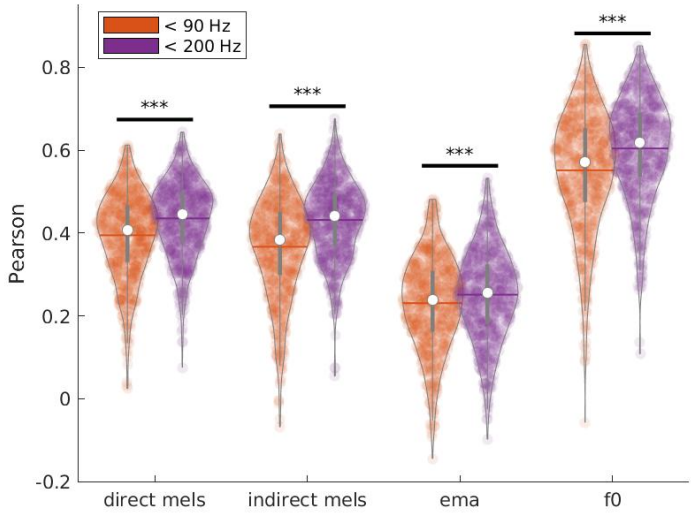
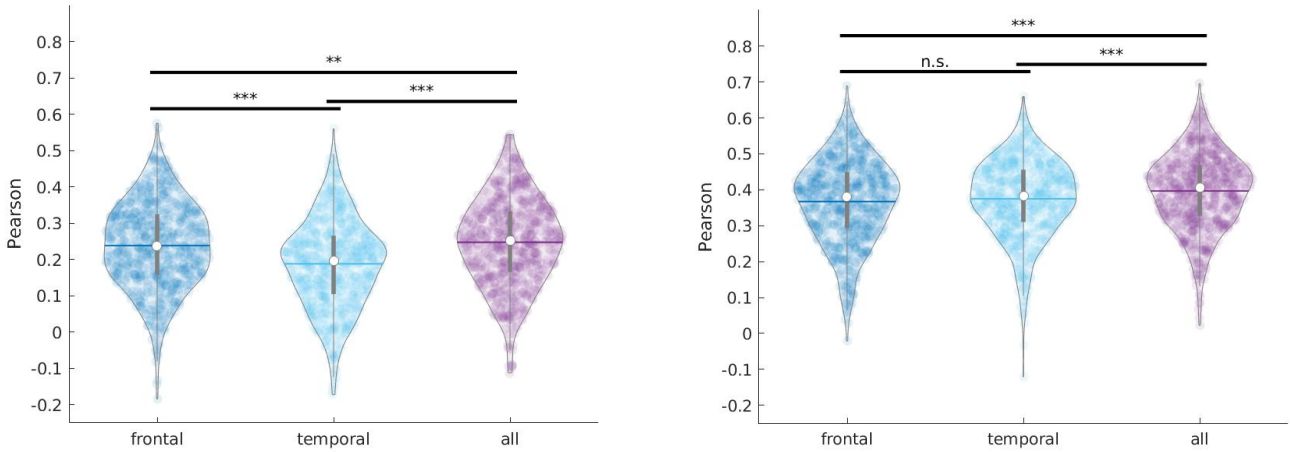


Fig. 4.23.: Influence of high gamma neural features on decoding of speech from P5 dataset. Correlations between decoded and ground truth mel cepstral coefficients of P5 are represented by violin plots. Mel cepstral coefficients decoded from neural features including high gamma frequencies up to 200 Hz (right, purple) are compared with the corresponding acoustic feature decoded from neural features up to 90 Hz (left, orange).
Wilcoxon signed rank test significance. ***: p<0.001

4.4.4.2. Frontal and temporal activity

Articulatory trajectories were decoded from P5 neural features from only frontal or temporal electrodes according to the delimitation shown in section 3.2.2.2. A PLS regression with 12 components was trained in each case and compared to the case where all electrodes were considered. Mel cepstral coefficients were then predicted by the articulatory-to-acoustic DNN with finetuning for each case.

Pearson correlations of decoded with ground truth features are displayed in figure 4.24 for P5 dataset, statistical significance was computed using Wilcoxon signed-rank tests. Several results were found. First, like for direct decoding of acoustic features (see 4.2.6), decoding with all electrodes increased correlations compared to isolated frontal or temporal electrodes for both articulatory trajectories ($p = 0.0028$ and $p = 5.8e^{-5}$, respectively) and indirect decoding of mel cepstral coefficients ($p = 0.0028$ and $p = 5.8e^{-5}$, respectively). Second, indirect decoding of mel cepstral coefficients was very similar with both frontal and temporal electrodes ($p = 0.06$, no significant difference), despite better decoding of articulators with frontal electrodes.



(a) Articulatory features

(b) Mel cepstrum: indirect decoding

Fig. 4.24.: Comparison of frontal and temporal electrodes for decoding of speech from P5 dataset. (a) Pearson correlations of decoded articulatory trajectories by a PLS regressions with 12 components from either frontal, temporal or all electrodes (median correlations r=0.24, r=0.20, r=0.25 for frontal, temporal and all electrodes respectively). (b) Indirect decoding of mel cepstral coefficients using an acoustic-to-articulatory DNN (r=0.38, r=0.38, r=0.41).
Wilcoxon signed rank test significance - ***: p<0.001, **: p<0.01

4.5 Evaluation of other neural features

4.5.1 Neural signals referencing

Acoustic and articulatory features of speech were decoded from neural features of P5 dataset after either **common median reference** or **bipolar reference**, which are two possible methods for removing artefacts from recorded signals (see section 3.2.1). Decoding was performed by a PLS regression with 12 components, combined with the finetuned articulatory-to-acoustic DNN for indirect decoding of mel cepstral coefficients.

The correlations of decoded ground truth features comparing both reference methods are displayed in figure 4.25. All features decoded from neural features with common median reference exhibited higher correlations than with bipolar reference.

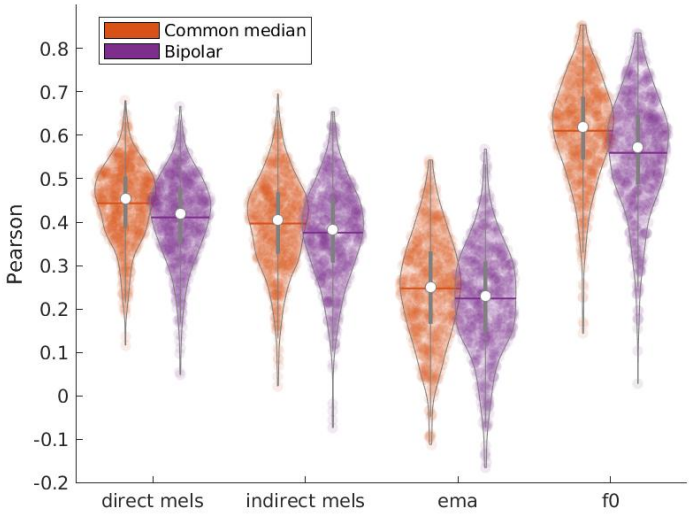


Fig. 4.25.: Comparison of bipolar *vs.* common median reference methods on speech decoding in P5 dataset. A PLS regression with 12 components was trained to predict mel cepstral coefficients (*direct mels*), articulatory trajectories (*ema*) and F0 from neural features of P5 dataset. A DNN was finetuned to predict mel cepstral coefficients from decoded articulatory trajectories (*indirect mels*). Violin plots display the correlations of decoded features with dataset's ground truth after **common median reference** (orange, left) or **bipolar reference** (purple, right). Median correlations are represented by white dots, while average correlations are shown by an horizontal line.

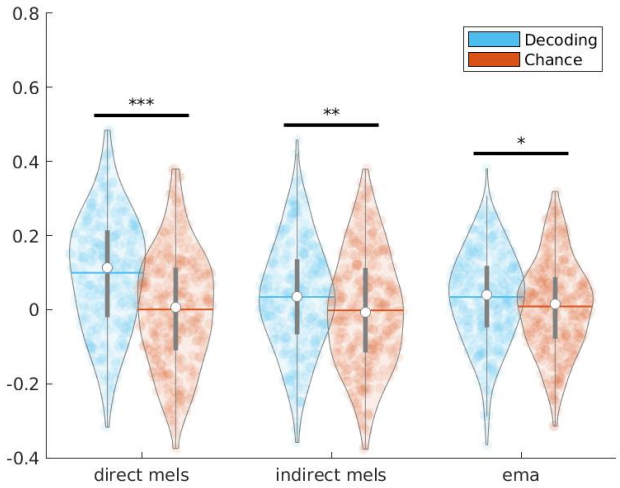


Fig. 4.26.: Decoding of speech from phase features of neural activity on P5 dataset. A linear regression with a PCA reduction of 100 components was trained to decode mel cepstral coefficients (*direct mels*), articulatory trajectories (*ema*), and a DNN was trained to predict mel cepstral coefficients from decoded articulatory trajectories (*indirect mels*). Each decoded features (left violin plots, blue) were compared to a chance level (right violin plots, red) using a Wilcoxon signed-rank test.
n.s.: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

4.5.2 Phases

A linear model was trained to predict mel cesptral coefficients and articulatory trajectories from phase information extracted from neural data (see 3.2.2.3). This was done only on P5 day 3 dataset. Phase features of neural activity were reduced to 100 PCA components and concatenated to create 10 frames of past and future contexts. The articulatory-to-acoustic DNN was finetuned to predict mel cepstral coefficients from decoded articulatory trajectories. A chance level was computed by shuffling neural features before running the exact same decoding pipelines.

All features were decoded above chance levels, albeit with low correlations (see figure 4.26). The highest median correlation was found for direct decoding of mel cepstral coefficients, with r=0.1 only.

4.6 Linear decoding of formants from neural features

P5 dataset's formants F1 and F2 and F0 (see 3.3.3.1) were predicted from 21 neural frequency bands (see 3.2.2.1) using a PLS regression with 12 components. Mohamed

Ben Ticha did the extraction of formants and I did the decoding from neural activity. Best neural features were selected by Welch t-test from overt sentences of days 1,2 and 3 of P5 dataset (1455 out of 1512 features, see table 4.3). Then, neural features were concatenated to form 10 frames of past and future context before training the PLS regression.

A PLS regression model was trained on complete sentences following the **continuous** regression paradigm, while another PLS regression was trained to predict voiced segments of speech only, following the **gated** paradigm. The description of both methods can be found in section 3.6.3.3. The decoding performances were evaluated in 10-fold crossvalidations by computing correlations and mean squared error of decoded features with the ground truth features extracted from P5 dataset.

4.6.1 Continuous regression

This section focuses on the evaluation of decoding of P5's formants and F0 using the continuous paradigm.

4.6.1.1. Decoding

Complete sentences of P5 were predicted from neural activity and evaluated by correlation and mean squared error. A chance level was computed by running the same decoding and evaluation pipeline on shuffled neural activity. Both decoding and chance levels are reported in figure 4.27, for both evaluation metrics.

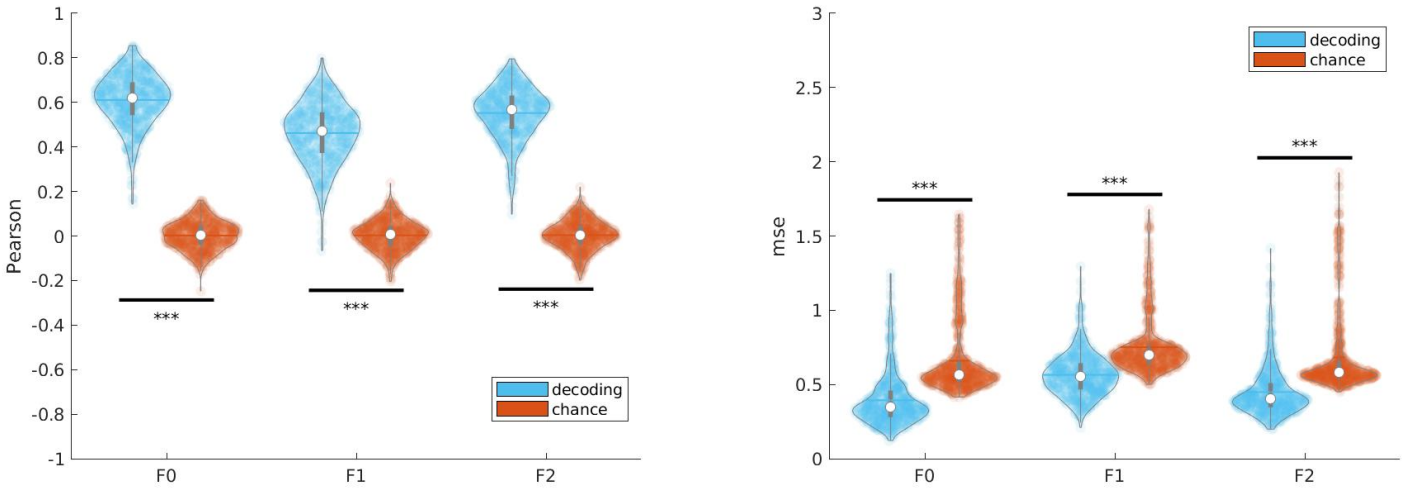
All F0, F1 and F2 show a significant decoding compared to chance levels ($p << 0.001$; Wilcoxon signed-rank test), both for correlation and MSE evaluations. Decoding of F0 performed better than formants while decoding of F2 was better than F1. The median scores of decoded sentences are reported in the tables below:

Correlations	*F0*	*F1*	*F2*
Decoded	0.62 ± 0.11	0.47 ± 0.14	0.57 ± 0.12
Chance	0.01 ± 0.06	0.01 ± 0.06	0.00 ± 0.06

MSE	*F0*	*F1*	*F2*
Decoded	0.35 ± 0.18	0.56 ± 0.14	0.41 ± 0.17
Chance	0.57 ± 0.25	0.70 ± 0.20	0.58 ± 0.27

4.6.1.2. Trajectories

An example of decoded formants and F0 trajectories was displayed on top of matching ground truth trajectories in figure 4.28. While the ground truth trajectories



(a) Correlations

(b) Mean squared error

Fig. 4.27.: Continuous PLS decoding of F0 and formants from P5 dataset. (a) Correlations of decoded formants and F0 with ground truth from P5 dataset. A chance level (right, red) was computed for each decoded feature (left, blue) by shuffling neural activity using a Wilcoxon signed-rank test. (b) Means squared error of decoded formants and F0 from P5 dataset (left, blue) and matching chance levels (right, red).
n.s.: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

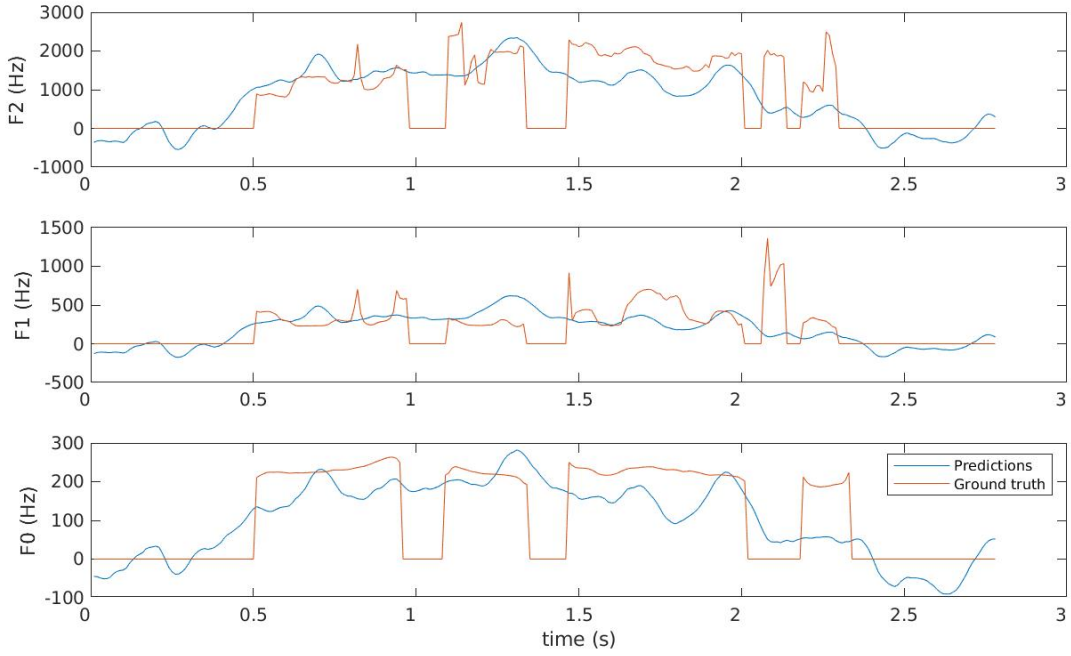


Fig. 4.28.: Formant trajectories of a decoded sentence of P5 with continuous regression: *'Au moins une sévère leçon.'*
From bottom to top: trajectories of decoded (blue) and ground truth (orange) F0, F1 and F2 in Hz, respectively. Time is represented on the horizontal axis in seconds at 100 Hz sampling rate.

includes discontinuities between voiced and voiceless segments (formants and F0 are set to 0 by convention on voiceless segments), the trajectories predicted by the regression do not reproduce those discontinuities. The regression approximates the general trend of the trajectories without accurately predicting trajectories on voiced segments.

4.6.2 Gated regression

This section compares the decoding of P5's formants and F0 using the gated paradigm and the continuous paradigm. In the gated paradigm, the regression model was trained on voiced segments only and its predictions are set to 0 outside of voiced segments using the dataset's annotation. This model is therefore trained to accurately predict voiced segments but should be combined with a classifier that detects voicing (see 3.6.3.3).

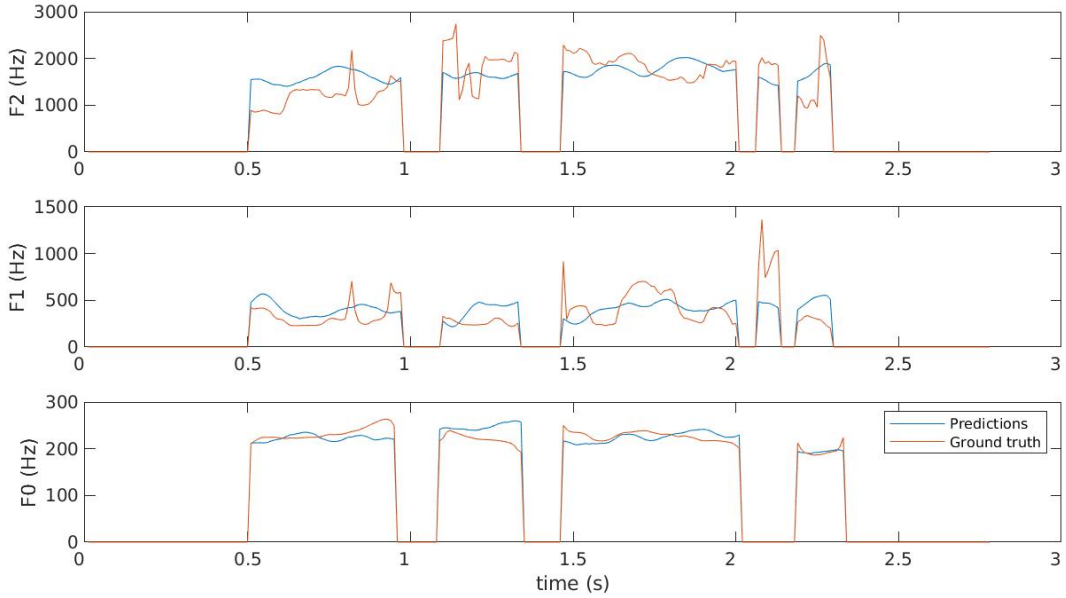


Fig. 4.29.: Formant trajectories of a decoded sentence of P5 with gated regression: *'Au moins une sévère leçon.'*
From bottom to top: trajectories of decoded (blue) and ground truth (orange) F0, F1 and F2, respectively. Time is represented on the horizontal axis in seconds at 100 Hz sampling rate.

4.6.2.1. Comparison of gated and continuous paradigms

Both gated and continuous paradigms of F0, F1 and F2 decoding were compared on P5 dataset. Decoded sentences from both paradigms were evaluated on both complete sentences and on voiced segments only. The correlations and mse are reported in figure 4.30 and the median scores of decoded sentences are reported in the tables below as well as p-values of Wilcoxon signed-rank test between continuous and gated scores:

Correlations	*F0 (all)*	*F0 (voiced)*	*F1 (all)*	*F1 (voiced)*	*F2 (all)*	*F2 (voiced)*
Continuous	0.62	0.21	0.47	0.02	0.57	-0.08
Gated	0.99	0.34	0.78	0.05	0.94	0.12
Significance	***	***	***	***	***	***
p-value	0	$5.0e^{-12}$	0	$8.8e^{-4}$	0	$3.1e^{-28}$

MSE	*F0 (all)*	*F0 (voiced)*	*F1 (all)*	*F1 (voiced)*	*F2 (all)*	*F2 (voiced)*
Continuous	0.35	0.16	0.56	0.34	0.41	0.22
Gated	0.01	0.01	0.30	0.30	0.07	0.07
Significance	***	***	***	***	***	***
p-value	0	0	0	$6.8e^{-31}$	0	$2.2e^{-94}$

Correlations and MSE improved for each features when using gated regressions compared to continuous regressions. The difference was larger when computing evaluation on complete sentences, however it remained consistently true when evaluating voiced segments only.

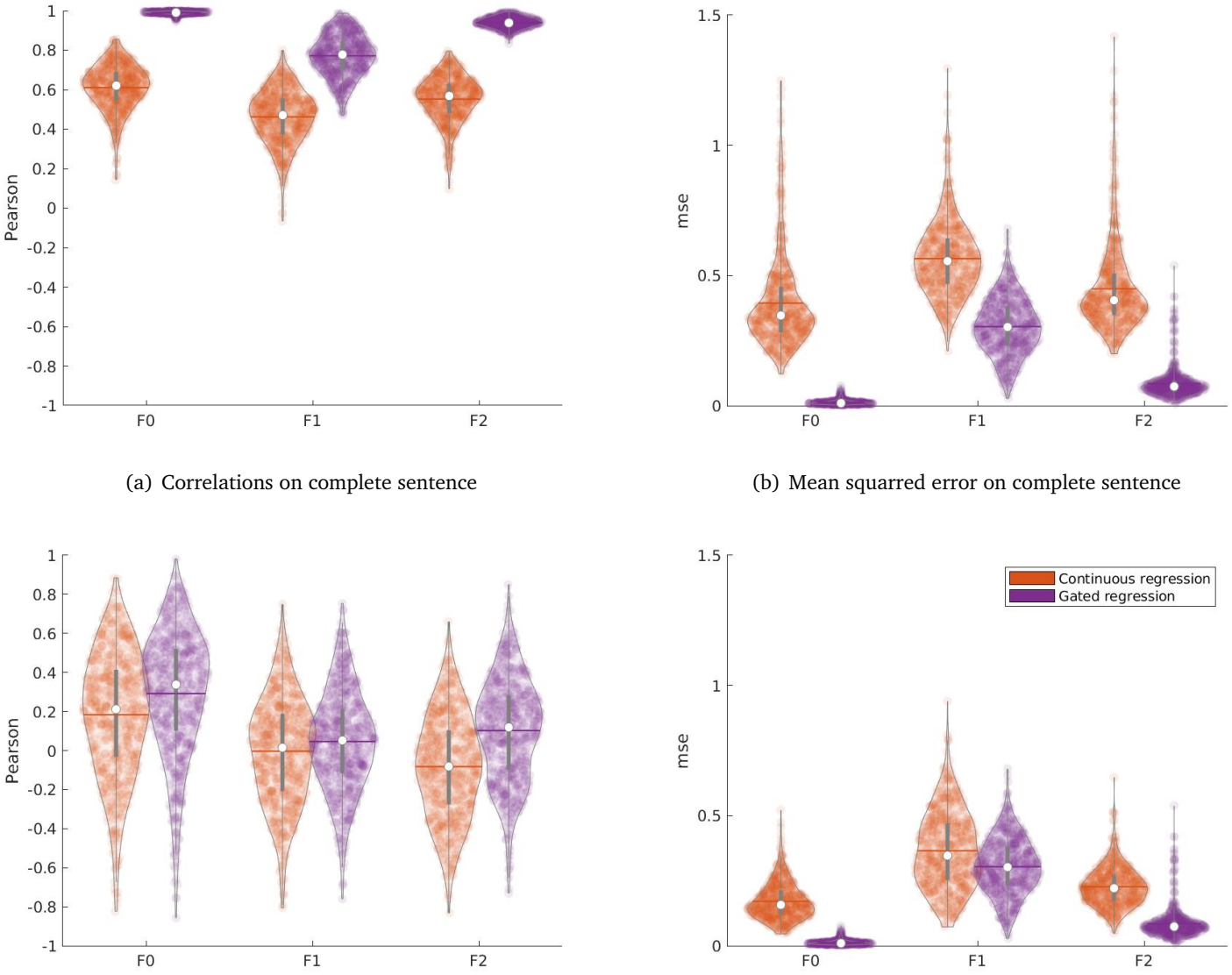
4.6.2.2. Decoding of voiced segments

Both gated and continuous decoding paradigms were compared to a chance level computed by randomly shuffling all neural features samples before training the PLS decoder. The correlations and MSE of decoded sentences with their corresponding ground truth were computed on voiced segments. The same was done to compute chance levels, on formants decoded from shuffled neural data. The evaluations of decoding and chance levels is reported in figure 4.31 for both gated and continuous paradigms.

Continuous decoding of F0 and formants shows better MSE than chance, however only F0 showed better correlations than chance. Reported F2 correlations were even below 0 and chance levels. On the other hand, correlations of F0, F1 and F2 decoded by gated regression were reported better than chance. MSE of decoded F0 and F2 were very close to chance levels, while MSE of decoded F1 were lower than chance levels.

4.6.2.3. Trajectories

The example trajectories decoded with the continuous paradigm in figure 4.28 were also displayed for gated decoding in 4.29. In this example, the theoretically perfect gating of course correctly reproduced the discontinuities of voicing, but it also allowed the decoder to better learn trajectories of voiced segments. On this particular example, F0 decoded trajectories are much closer in values than for continuous regression. It also appears to be the case for F2, although to a smaller extent.



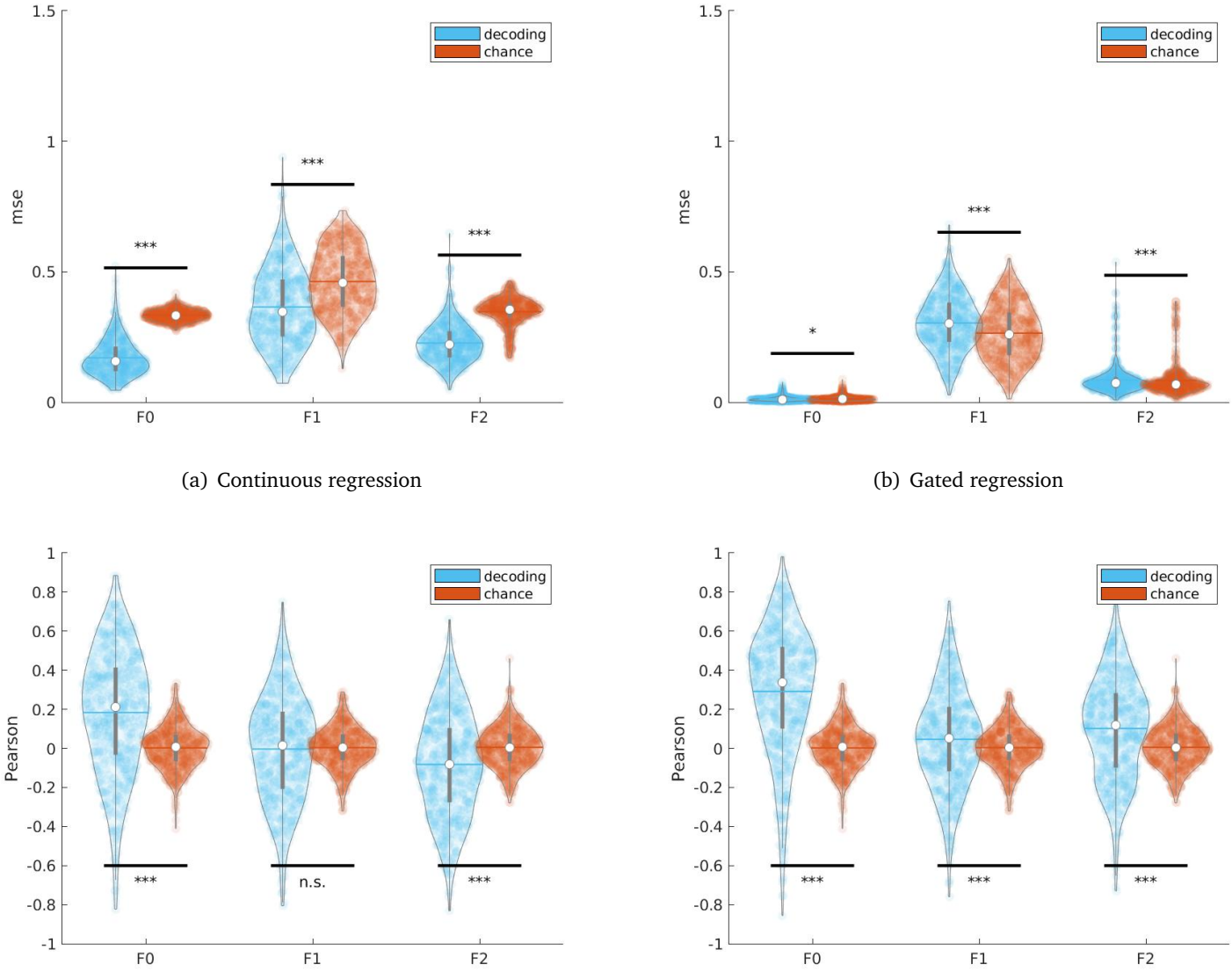
(a) Correlations on complete sentence

(b) Mean squarred error on complete sentence

(c) Correlations on voiced segments

(d) Mean squared error on voiced segments

Fig. 4.30.: Comparison of continuous and gated formant decoding from P5 dataset using PLS regression. (a) Correlations of decoded formants and F0 with ground truth computed on complete sentences. Decoding with continuous regression (left, orange) was compared with gated regression (right, purple). (b) Mean squared error of decoded formants and F0 on complete sentences for continuous (left, orange) and gated (right, purple) regressions. (c) Correlations of decoded formants and F0 on voiced segments for continuous (left, orange) and gated(right, purple) regressions. (d) Mean squared error of decoded formants and F0 on voiced segments for continuous (left, orange) and gated(right, purple) regressions.



(a) Continuous regression

(b) Gated regression

(c) Continuous regression

(d) Gated regression

Fig. 4.31.: Decoding formants on voiced segments: chance levels with and without gating. Decoded formants and F0 (left violin plots, blue) were evaluated on voiced segments as well as their matching chance levels (right violin plots, orange). Differences between decoding scores and chance levels were reported using a Wilcoxon signed-rank test. (a) Mean squared error of formants and F0 decoded by continuous regression. (b) Mean squared error of formants and F0 decoded by gated regression. (c) Correlations of formants and F0 decoded by continuous regression. (d) Correlations of formants and F0 decoded by gated regression.
n.s.: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

4.7 Further work

4.7.1 Comparing linear and neural networks based methods for speech decoding

Xingchen Ran, a visiting PhD. student from the lab developed a neural network based decoding of speech from neural activity on P5 dataset. I adapted it to run on EC61 dataset and evaluated its predictions of mel cepstral coefficients and F0 compared to PLS regressions.

Decoding of mel cepstral coefficients and F0 were performed by PLS regressions with 18 and 12 components respectively, according to the best results found in section 4.2.4. The comparison of correlations of both decoding methods are shown in Fig. 4.32. Median correlations of decoded average mel cepstral coefficients achieved r=0.46 and r=0.36 for ANN and PLS respectively, while median correlations of decoded F0 achieved r=0.54 and r=0.51 for ANN and PLS respectively. Wilcoxon rank-sum tests showed that decoding correlations increased significantly for each acoustic features compared to PLS regression (all p-values below 0.001). A rank-sum test was used instead of sign-rank because both distributions were not paired: a few sentences were eliminated in the processing of EC before the training of the ANN.

Example spectrograms of a decoded sentence using both the PLS and ANN decoding is shown in Fig. 4.33. Decoded sentence using PLS regression was synthesized using an MLSA filter from F0 and mel cepstral coefficients, while decoded sentence using the ANN was synthesized using WORLD synthesizer from F0, mel cepstral coefficients and aperiodicity. On this example, the ANN shows a more precise decoding of voicing and a better reproduction of change of power in higher frequencies, which allows to distinguish consonants. However, neither methods properly decode formant trajectories that are clearly visible in the original sentence. Although WORLD synthesis is admittedly better than MLSA synthesis, most of the differences between synthesized sentences are explained by the improvement of decoding of F0, voicing, and mel cepstral coefficients when using the ANN.

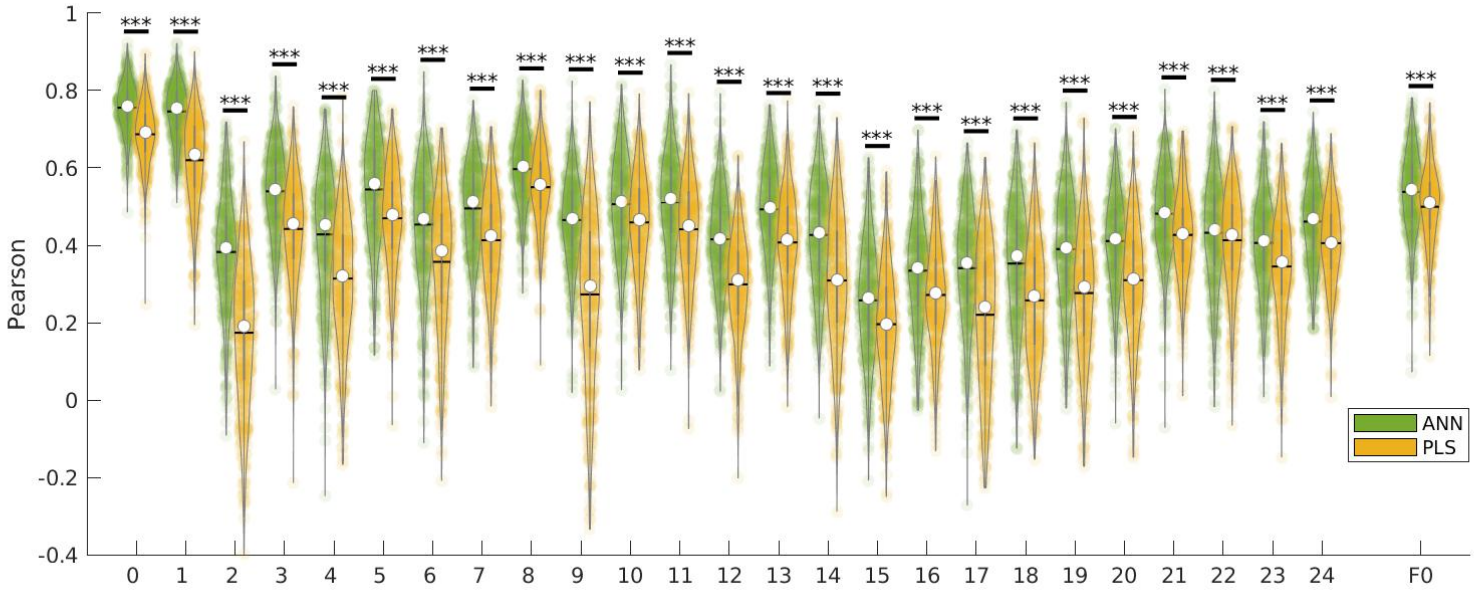
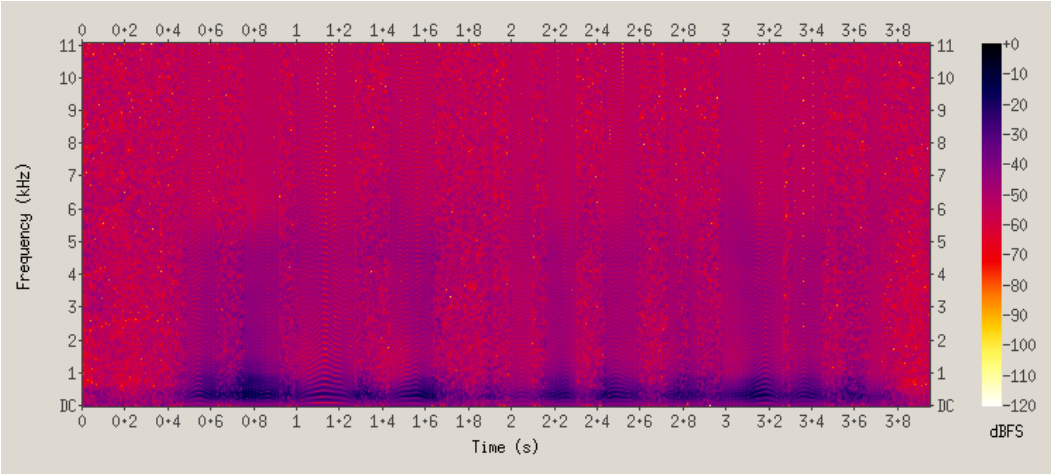
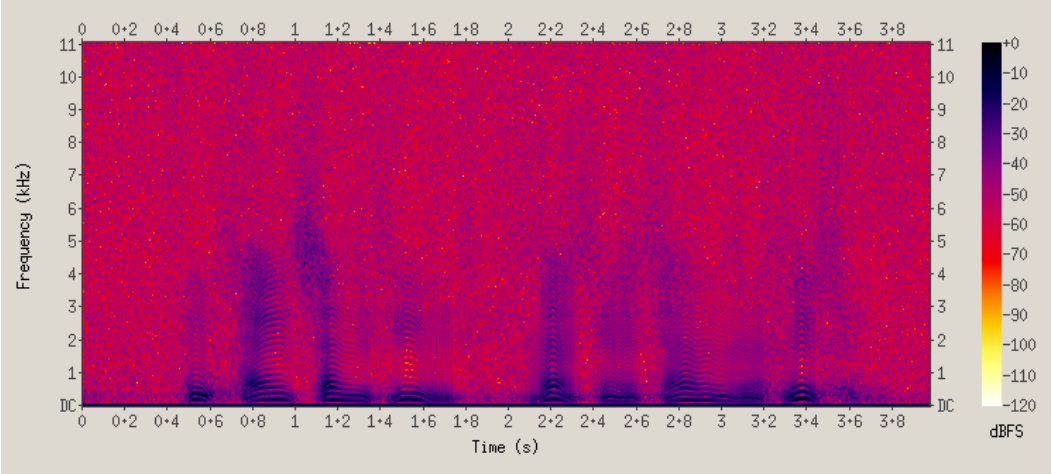


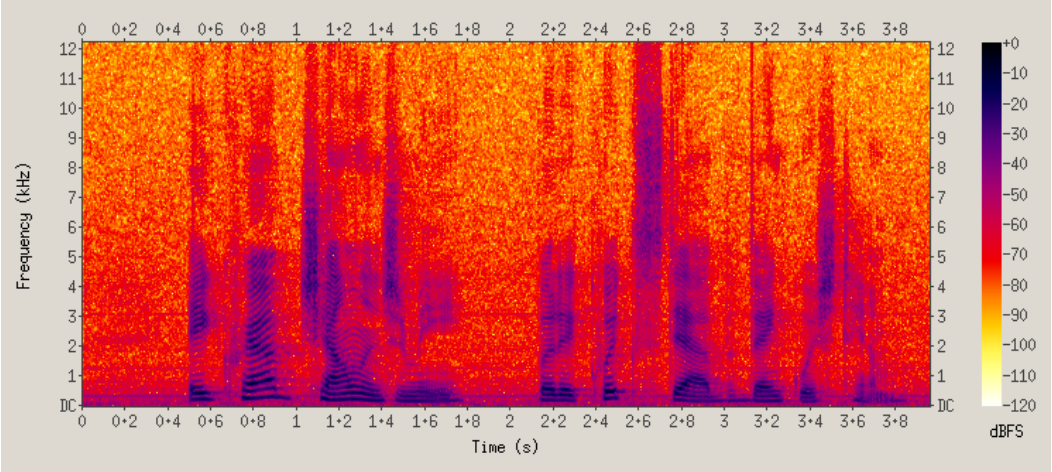
Fig. 4.32.: Comparison of acoustic features decoding using ANN and PLS on EC61 dataset. Correlations of decoded mel cepstral coefficients (0 to 24) and F0 from EC61's neural activity using artificial neural networks (green, left) and PLS regressions (yellow, right). The PLS regression trained to decode mel cepstral coefficients used 12 components, while the one trained to decode F0 used 18 components. Statistical differences were assessed by a Wilcoxon rank-sum test: ***: $p < 0.001$



(a) Decoded with PLS



(b) Decoded with ANN



(c) Original

Fig. 4.33.: Spectrograms of decoded sentence from EC61 dataset: *'A crab challenged me but a quick stab vanquished him.'* (a) MLSA synthesis of decoded F0 and mel cepstral coefficients from cortical activity using a PLS with 18 and 12 components, respectively. (b) WORLD synthesis of decoded F0, mel cepstral coefficients and aperiodicity from cortical activity using a neural network. (c) Original sentence recorded in EC61 dataset.

5 General conclusion

5.1 Neural features for speech decoding

5.1.1 Feature Selection

Automatic feature selection using Welch's *t*-test with Bonferroni correction successfully removed noisy channels that did not properly record brain activity on P2's dataset (see 4.1.1). Indeed, as P2's ECoG grid was too rigid to closely fit the cortex, multiple channels were not in contact with the cortex and thus recorded only noise. This indicates that automatic feature selection could replace visual inspection of each channels.

Decoding of mel cepstral coefficients from P2's neural activity was significantly increased after removing noisy channels (see 4.2.1). Feature selection seemingly removed some polluting information from P2 before PCA reduction. When performing a PCA to reduce the number of features prior to training the linear model, there is indeed no mechanism that prioritizes speech related neural features compared to any other neural activity or even noise. Performing a feature selection before the PCA might improve the speech related information of the first PCA components. On P5 dataset on the other hand, decoding of mel cepstral coefficients from neural activity was only slightly impacted by feature selection (see 4.2.1). This suggests that feature selection indeed removed features irrelevant for decoding from P5 dataset, but also that PCA already selects features with speech information. In the case of P5, feature selection seems to not be critical for decoding when using feature reduction like PCA or PLS.

Other than potentially improving decoding, feature selection would also improve the memory cost of speech decoding algorithms. In practice that was true for P2 and individual days of P5 for which a large part of the channels were dismissed, but not for the complete P5 dataset combining days 1,2 and 3 as most features remained. The large amount of dismissed features on P2 can be explained by the removal of noisy channels. For P5 however, the selection changed across the different days of recording and never entirely removed channels. All channels show some significant speech activity all across the temporal cortex, which is coherent with the current

knowledge of underlying cortical processes of speech production and perception, which are known to be widely spread across the temporal cortex, premotor cortex and sensorimotor cortex (Hickok and Poeppel, 2007; Tourville and Guenther, 2011). The selection of different features depending on the day might be explained by a fluctuation of neural activity in between sessions, thus changing which features are modulated by speech activity. This would also partly explain why most features are selected when grouping the days along with the increasing amount of data: by aggregating more data, more features get modulated by speech activity at one point or another, therefore increasing the number of selected features.

5.1.2 Neural data cleaning

Common median referencing was found to increase decoding of articulatory and acoustic speech features compared to bipolar referencing, for both direct and indirect decoding (see section 4.5.1). Hence why we decided to use Common Median Referencing for all other experiments. We did not test it however on EC61 dataset, for which another group used bipolar referencing (Anumanchipalli et al., 2019).

5.1.3 Relevant neural features for speech decoding

5.1.3.1. Acoustic contamination

Our results show that the contribution of high gammas to speech decoding is higher for P2 than for P5, which is consistent with our expectations that speech decoding would be improved by acoustic contamination in the high gamma range (see 4.2.2). However it does not establish a causality link.

Therefore, P2's dataset and other datasets that exhibit acoustic contamination below 200 Hz should only be used for decoding after removing spectral content above the lowest F0 values of the patient from neural activity (90Hz for P2). P5 on the contrary proved to be free of acoustic contamination below 200Hz as P5's high pitched female voice did not cover lower frequencies. As P2 showed acoustic contamination as well as poorer decoding compared to P5 and less data than the complete P5 dataset, the next experiments focused on P5 only.

5.1.3.2. Optimal temporal time span

Decoding of acoustic and articulatory features of speech from neural activity on P5 dataset was shown to increase with temporal context size, up to 210 ms (see 4.2.5.1, 4.4.1.3). While it might still increase with more context, we could not test it on the whole dataset as it was maxing out the RAM of our computing server. However, on prior experiments on a smaller subset of P5 dataset, we found that increasing time context up to 310 ms actually decreased decoding correlations compared to 210 ms context.

In addition to a 210 ms time context, we found in sections 4.2.5.2 and 4.4.1.3 that decoding speech from neural activity happening between 200 to 0 ms prior was optimal (Fig. 4.11). This result tends to show that speech was actually decoded from neural activity related to speech intent more than auditory and sensory feedback. On a practical side, the best decoding model would be able to run in real-time for a closed-loop speech BCI.

5.1.3.3. Spatial organization

Cortical maps of the p-values computed for feature selection were shown for each frequency band in section 4.1.2. The statistical meaning of the *p-value* alone cannot be rigorously interpreted as a score of how much a neural feature is modulated by speech. It does not even prove that features with a *p-value* higher than 0.05 do not actually hold speech information. However the results from section 4.2.1 tend to show that discarded features do not display sufficient variance to contribute significantly to speech decoding after PCA, and that the selected features indeed hold speech information (see 5.1.1). The results show that all brain areas recorded by the ECoG grid show some speech related activity in at least one frequency band, including both frontal and temporal electrodes. It should be noted that the mapping results were computed from a single day of recording. A mapping based on another day of recording might show a different spatial organization of speech activity, and a mapping based on all days of recording would show speech activity in almost every single feature.

Decoding acoustic features of speech from temporal electrodes yielded slightly but significantly higher correlations than from frontal electrodes (see 4.2.6). On the contrary, decoding of articulatory features of speech from frontal electrodes yielded higher correlations than from temporal electrodes (see 4.4.4.2). That would be coherent with the fact the cortical activity related to speech articulators

is mainly found in the frontal areas, while activity related to acoustic processing is predominantly found in temporal areas.

For all speech features however, using all electrodes for speech decoding performed significantly better than using only frontal or temporal electrodes. That means that frontal and temporal electrodes contain at least some non-overlapping information about either form of speech representation, which further supports the current understanding of cortical mechanisms of speech as distributed specialized cortical processes across the frontal and temporal lobes (Hickok and Poeppel, 2007; Tourville and Guenther, 2011). A speech BCI should therefore use all electrodes, no matter which features of speech it decodes from neural activity.

5.1.3.4. Spectral organization

Cortical maps of the p-values computed for feature selection show significant speech related neural activity over all frequency bands, including high gamma frequencies 4.1.2. Including high gamma frequencies above 90Hz for decoding of speech increased correlations for all acoustic and articulatory speech features (see section 4.4.4.1), which is consistent with the literature where high gammas have been consistently found to contribute to speech decoding (Leuthardt et al., 2004; Bouchard et al., 2013; Moses et al., 2018; Anumanchipalli et al., 2019). However decoding without high gammas still performs fairly well, meaning that other frequency bands largely contribute to decoding. The p-value maps show relevant speech activity below 90Hz.

Other speech decoding studies have already been including both high gammas and lower frequency bands to neural features (Ibayashi et al., 2018; Anumanchipalli et al., 2019) and adding lower frequencies was found to be beneficial (Akbari et al., 2019). Lower and higher frequency bands are indeed believed to encode different and complementary activities (Buzsáki et al., 2012). Further work should investigate contribution of individual features to decoding by looking at the weights of linear decoding methods. That would give more precise insight on which frequency bands influence decoding the most as well as their spatial location and time delay with actual sound production.

Aside from spectral amplitude features, preliminary work showed that phase features include relevant information for speech decoding although decoding correlations are low (see section 4.5.2). Further work should evaluate if phase features provide complementary information to spectral amplitude features of speech, which might improve decoding.

5.2 Speech decoding using vocoder-based synthesis

5.2.1 Direct decoding

5.2.1.1. Feature reduction

Our results show that PCA-based linear decoding of acoustic features of speech improves with the number of PCA components (see 4.2.3.1), with a maximum at 200 components. We did not test more features, as the memory of our computing server was maxed out. In these experiments, the PCA was computed before concatenating frames for temporal context, as a preliminary experiment showed that computing PCA after temporal context decreased decoding correlations.

On the other hand, feature reduction using PLS showed best correlation for 12 components on P5 dataset (see 4.2.3.2). This cannot be directly compared with the PCA results, as the PLS reduction was computed after concatenating frames for temporal context. However, in order to assess the best feature representation for decoding, the PLS reduction shows a much more compact representation for similar decoding performance.

On EC61 dataset, the optimal number of PLS components was shown to be similar with 12 and 18 components for mel cepstral coefficients and F0 decoding, respectively. That indicates that 12 PLS components would be a good estimate for building a speech BCI. Although the number of ECoG channels is larger in EC61 than in P5, those similar numbers could be mitigated by the selection of the 1500 best features of EC61, which is similar to the number of selected features of P5.

5.2.1.2. Decoding methods

Decoding of articulatory and acoustic features of speech performed similarly for all linear regression methods (see 4.4.1.1 and 4.2.3.3, respectively). Those results tend to invalidate the necessity for ridge regularization of the linear regression on top of the PCA. While ridge regularization adds a significant processing overhead, whichever method was used to optimize the λ factor, it did not provide any significant decoding improvement over a basic linear regression with PCA. That could be explained by the size of P5 dataset, maybe regularization would be more useful on smaller datasets. Those results also show again that PLS regression performs as well as other PCA-based linear regressions for a much more compact feature

representation: 12 components instead of 2100 components in this case (100 PCA components * 21 contextual frames).

5.2.2 Indirect decoding

5.2.2.1. Articulatory-to-acoustic synthesizer

We investigated the design of a neural networks-based articulatory-to-acoustic synthesizer using variable temporal contexts. Although one could expect that no temporal context is needed for this task as one shape of the vocal track should always produce the same sound, temporal context is typically found in the literature to improve articulatory-to-acoustic synthesis (Aryal and Gutierrez-Osuna, 2016; Liu et al., 2016; Gosztolya et al., 2019). Our results corroborated this claim using both objective (see 4.3.1.1) and perceptive evaluation (see 4.3.2). A possible explanation could be that temporal context would provide information about the vocal tract that is not provided by EMA recordings, which only samples a few points of the vocal track.

We found however that using short temporal context with a feedforward DNN outperformed using arbitrary long temporal context with a recurrent neural network, at least for the objective evaluation. This result is not as clear with the perceptive evaluation as both methods performed similarly for normal sentences, a difference was only found for VCVs. Other studies previously found that recurrent neural networks, and especially bidirectional LSTMs, improved articulatory-to-acoustic synthesis over feedforward DNNs without temporal context (Liu et al., 2016; Taguchi and Kaburagi, 2018; Cao et al., 2018). But those were not compared to DNNs with finite temporal context, even though temporal context was found to improve articulatory-to-acoustic synthesis with both feedforward DNNs (Aryal and Gutierrez-Osuna, 2016) and autoencoders (Gosztolya et al., 2019). The theoretically unnecessary temporal context might partly explain why feedforward DNNs manage to outperform a bidirectional GRU. Another possible explanation could be the discrepancy in amount of data available to train neural networks due to the mere architecture of the networks: for the training of the GRU, there are as many data samples as sentences, while for DNNs there are almost as many data samples as frames in the dataset.

Despite some significant differences in perceptive evaluation of the various methods, the test suffers from a few limitations. First, even for a MUSHRA test the number of participants is small, and they exhibit a wide range of rating behaviours, some rating systematically lower, some higher, some with a large variance. Second the test

was limited to 9 VCVs and 10 sentences in order to keep it under 30 minutes. That could be an issue as neural network predictions may vary greatly from one sentence to another. Third, from a subjective point of view all models perform decently at a similar level, which would imply the need for a higher number of participants and items. It is indeed fairly difficult to distinguish the different models by ear, even with training.

Nevertheless, our results indicate that a real-time compatible articulatory-to-acoustic DNN could be used for a closed-loop BCI. According to our results, such a DNN should use 10 frames of past temporal context or more, as well as a limited future context (see 4.3.1.1). The latency induced by the best model on BY2014 is minimal with a single frame of future context *i.e.* 5ms (see 4.3.1.2). In a context of real-time use of the articulatory-to-speech DNN as a speech synthesizer, such a low latency should not cause any issue to be controlled in a closed-loop setting such as described by Bocquelet (2017).

In order to train an articulatory-to-speech synthesizer for indirect decoding of speech from neural activity, one needs to first estimate the articulatory trajectories of the patient. Our approach was to use dynamic time warping on an EMA dataset that contains the same sentences pronounced by the patient. Other studies, chose instead to train a neural network on multiple datasets to perform speaker-independent acoustic-to-articulatory inversion (Chartier et al., 2018; Anumanchipalli et al., 2019). This approach would theoretically allow to predict articulatory trajectories of a brand new sentence never seen in the training dataset spoken by a new speaker, while ours requires to use a specific set of sentences for calibration. On the other hand, our method only requires a reference and a target sentence and works without any training, while training a neural network for acoustic-to-articulatory inversion requires a lot of data. Both methods seem to perform a very good estimation of articulatory trajectories, although we did not rigorously evaluate ours other than listening to the aligned sentences.

A potential direction for further work would be to improve the control parameters of the speech synthesizer. The articulatory features obtained by EMA recording have redundant information. Previous work already investigated reduction of the articulatory features using naive PCA and deep autoencoders, showing that deep autoencoders were more efficients (Bocquelet et al., 2014). It would also be possible to use a guided PCA so that only a few orthogonal and meaningful articulatory features are left (Bailly et al., 2006). This method consists in assisting the PCA with knowledge about the articulators, for example by imposing components for the jaw and then removing the jaw trajectories from the tongue before computing a PCA

over the tongue trajectories only. Such features could reduce the degrees of freedom of the articulatory-to-acoustic synthesizer, while still offering natural and therefore possibly easier control for a speech BCI.

5.2.2.2. Comparison of decoding methods

All linear methods used for articulatory trajectories decoding showed similar performance (see 4.4.1), showing again that PLS regression is probably the better linear decoding method as we already discussed about acoustic features in section 5.2.1.2.

Compared to acoustic features, decoding correlations of articulatory trajectories were lower. Investigating decoding correlations of individual articulatory trajectories using PLS regression show large discrepancies (see 4.17). The velum is very well decoded, which could likely be attributed to its predictable trajectories: during sentence production, it most often lifts up to close the nasal cavity when starting to speak, and opens again at the end of the sentence to breathe through the nose. The best decoded trajectory after the velum is the vertical position of the lower lip, which is especially important to predict lips closure. The tongue position is poorly decoded, especially on its horizontal axis. That could be explained by P5's limited coverage of the sensorimotor cortex, considering that the tongue was found to be represented on the dorsal part of the ventral sensorimotor cortex Bouchard et al. (2013). A better decoding of the tongue would be essential to correctly predict the tongue constrictions that characterize consonants and vowels, combined with the already good decoding of the velum and lips closure. Vowels in particular would benefit from a better decoding of the horizontal trajectories of the tongue and lips. Consonants would probably benefit more from a better decoding of the vertical movements of the tongue, as they are mostly characterized by the place of articulation *i.e.* which part of the tongue is raised, if any.

Despite the poor decoding of articulatory trajectories, indirect decoding of acoustic features performed similarly to direct decoding, with exception of indirect decoding using PLS regression for which correlations were significantly lower than direct decoding correlations (see 4.4.3). The lower performance using PLS regression is unexpected, considering that the decoding performance of articulatory trajectories was evaluated to be similar using all linear methods. Indirect decoding was only tested on P5 dataset, although it would also be possible to test it on EC61 dataset after estimating its articulatory trajectories from MOCHA dataset.

This good performance of indirect decoding relied on transfer learning, without which decoding correlations largely decrease (see 4.4.2). After finetuning on a training set of decoded articulatory trajectories, the articulatory-to-acoustic DNN manages to reconstruct better acoustic features. This does not show however if the DNN used previous information from its training on BY2014, further work should also evaluate a DNN trained from scratch on the training set of decoded articulatory trajectories. Furthermore, the DNN already proved to be more efficient than linear methods on prediction of acoustic features from proper articulatory trajectories, but we did not compare both methods on decoded articulatory trajectories which are noisier.

5.3 Speech decoding using formant-based synthesis

Our results show that PLS regression can be successfully used to decode formants and F0 (see 4.6.1.1), using the same method that was used to decode F0 for vocoder-based speech synthesis. However we found that such naive approach could not properly reproduce the discontinuities between voiced and voiceless segments (see 4.6.1.2). We introduced a gating mechanism that separates the detection of voicing and the prediction of F0 and formants. However our work did not include a voicing decoder yet, and only evaluated a PLS regression trained to decode voiced segments as if we had a perfect voicing decoder.

Gating proved to largely improve F0 and formants decoding on complete sentences compared to continuous decoded but this comparison is unfair, as gated decoding used the ground truth detected voicing. However the improvement of decoding on voiced segments shows that gating improves the training of the regression on voiced segments (see 4.6.2.1). It not only improved correlations, which hints at a better decoding of formants transitions, but also MSE, which implies a better decoding of the formant absolute values. This is especially important as the intelligibility of speech is determined by the formants absolute values. On the other hand, the relevant semantic information conveyed by the voice's F0 is mostly determined by its trajectories which are better evaluated by correlations: the absolute value of the F0 mostly influences how low the voice is perceived, while its trajectories define the prosody and intonation of speech.

Both gated and continuous decoding paradigms were compared to chance levels on voiced segments (see 4.6.2.2). Formants decoded by the continuous paradigm showed poor and even negative correlations, while the MSE were significantly better

than chance. The MSE implies that the regression significantly predicts the switch between voiced and voiceless segments, however it does not properly decode the trajectories on the voiced segments. Although correlations of F0, F1 and F2 decoded by gated regression showed an improvement compared to continuous regressions, their MSE are close to chance levels which indicates that the gating mechanism itself largely improves decoding. MSE is an important metric to consider for formants as their absolute values better describe intelligibility of speech than their trend.

A speech decoder based on formant synthesis could benefit from using a gated regression paradigm. Our experiments show that good prediction of the voicing is critical to the overall performance of the gated regression. Further work should therefore investigate how to build a powerful classifier of voicing from neural signals.

5.4 Speech decoding using neural networks

This manuscript focused on linear methods for speech decoding. However further work should investigate neural network-based approaches, which already proved to achieve state of the art decoding in the literature. We presented in section 4.7.1 some preliminary work using an artificial neural network that improved decoding correlations on EC61 dataset compared to linear methods. The resulting synthesized sounds were subjectively closer to intelligibility although not intelligible yet on EC61. This decoding model is however not real-time compatible as it processes complete sentences at once, which is also the case of the continuous speech decoder described by Anumanchipalli et al. (2019). Further work should focus on adapting its architecture to be real-time compatible.

5.5 Perspectives

The work presented in this thesis paved the way towards a natural speech BCI by evaluating linear methods for speech decoding from invasive cortical recordings. In particular, partial least squares regression was found to reach competitive performance compared to other linear methods while largely reducing the number of decoding features. An indirect decoding method through an articulatory representation was also proposed, using dynamic time warping to estimate a dataset of articulatory trajectories and an efficient real-time compatible neural network for

articulatory-to-acoustic synthesis. It is still open whether one method would perform better than the other in a close-loop BCI setting, further work should therefore still investigate decoding through an articulatory representation. Additionally, feature selection using a Welch t-test was found to automatically remove noisy channels and neural features without speech information. Both direct and indirect decoding methods are real-time compatible, and should now be implemented for real-time BCI use in the lab's software PulsIO.

Decoding of formants and F0 from neural activity would be improved by a combined decoding model using both a regression and a voicing classifier. Further work should now focus on designing a good voicing classifier, as it was found to be crucial for decoding performance. Furthermore, this method could be extended to the decoding of mel cepstral coefficients, by combining the regression methods with a speech classifier.

Finally, preliminary results showed the potential of neural network based approaches for speech decoding from cortical activity. Those methods should further be studied while ensuring their design is real-time compatible. They should also combine classification and regression models to improve decoding, and might also benefit from including articulatory representations.

Bibliography

Abdoun, Oussama, Sébastien Joucla, Claire Mazzocco, and Blaise Yvert (2011). "NeuroMap: A Spline-Based Interactive Open-Source Software for Spatiotemporal Mapping of 2D and 3D MEA Data". In: *Frontiers in Neuroinformatics* 4, p. 119 (cit. on p. 103).

Ahmed, N, T Natarajan, and K R Rao (1974). "Discrete Cosine Transfonn". In: *IEEE TRANSACTIONS ON COMPUTERS*, p. 4 (cit. on p. 191).

Ajiboye, A Bolu, Francis R Willett, Daniel R Young, William D Memberg, Brian A Murphy, et al. (May 6, 2017). "Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration". In: *The Lancet* 389.10081, pp. 1821–1830 (cit. on pp. 50, 51).

Akbari, Hassan, Bahar Khalighinejad, Jose L. Herrero, Ashesh D. Mehta, and Nima Mesgarani (Dec. 2019). "Towards reconstructing intelligible speech from the human auditory cortex". In: *Scientific Reports* 9.1 (cit. on pp. 58, 148).

Allison, Brendan Z., Dennis J. McFarland, Gerwin Schalk, Shi Dong Zheng, Melody Moore Jackson, and Jonathan R. Wolpaw (Feb. 1, 2008). "Towards an independent brain–computer interface using steady state visual evoked potentials". In: *Clinical Neurophysiology* 119.2, pp. 399–408 (cit. on p. 45).

Angrick, Miguel, Christian Herff, Garett Johnson, Jerry Shih, Dean Krusienski, and Tanja Schultz (2018). "Interpretation of Convolutional Neural Networks for Speech Regression from Electrocorticography". In: *ESANN*, p. 6 (cit. on p. 61).

Angrick, Miguel, Maarten C. Ottenhoff, Lorenz Diener, Darius Ivucic, Gabriel Ivucic, et al. (Sept. 23, 2021). "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity". In: *Communications Biology* 4.1, pp. 1–10 (cit. on p. 54).

Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang (Apr. 2019). "Speech synthesis from neural decoding of spoken sentences". In: *Nature* 568.7753, pp. 493–498 (cit. on pp. 56, 59, 61, 69, 146, 148, 151, 154).

Arai, Takayuki (2004). "HISTORY OF CHIBA AND KAJIYAMA AND THEIR INFLUENCE IN MODERN SPEECH SCIENCE". In: p. 6 (cit. on pp. 16, 17).

Aryal, Sandesh and Ricardo Gutierrez-Osuna (Mar. 2016). "Data driven articulatory synthesis with deep neural networks". In: *Computer Speech & Language* 36, pp. 260–273 (cit. on pp. 82, 150).

Bailey, Dale L, Michael N Maisey, David W Townsend, and Peter E Valk (2005). *Positron emission tomography*. Springer (cit. on p. 31).

Bailly, Gérard, Frédéric Elisei, Pierre Badin, and Christophe Savariaux (2006). “Degrees of freedom of facial movements in face-to-face conversational speech”. In: *International Workshop on Multimodal Corpora*. Genoa, Italy, pp. 33–36 (cit. on p. 151).

Bartels, Jess, Dinal Andreasen, Princewill Ehirim, Hui Mao, Steven Seibert, E. Joe Wright, and Philip Kennedy (Sept. 30, 2008). “Neurotrophic electrode: Method of assembly and implantation into human motor speech cortex”. In: *Journal of Neuroscience Methods* 174.2, pp. 168–176 (cit. on p. 59).

Baykara, E., C. A. Ruf, C. Fioravanti, I. Käthner, N. Simon, S. C. Kleih, A. Kübler, and S. Halder (Jan. 1, 2016). “Effects of training and motivation on auditory P300 brain–computer interface performance”. In: *Clinical Neurophysiology* 127.1, pp. 379–387 (cit. on p. 48).

Beisteiner, R., P. Höllinger, G. Lindinger, W. Lang, and A. Berthoz (Mar. 1, 1995). “Mental representations of movements. Brain potentials associated with imagination of hand movements”. In: *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 96.2, pp. 183–193 (cit. on p. 45).

Benabid, Alim Louis, Thomas Costecalde, Andrey Eliseyev, Guillaume Charvet, Alexandre Verney, et al. (Dec. 2019). “An exoskeleton controlled by an epidural wireless brain–machine interface in a tetraplegic patient: a proof-of-concept demonstration”. In: *The Lancet Neurology* 18.12, pp. 1112–1122 (cit. on p. 50).

Berger, Hans (1929). “Über das elektroenkephalogramm des menschen”. In: *Archiv für psychiatrie und nervenkrankheiten* 87.1, pp. 527–570 (cit. on pp. 31, 38, 46).

Bin, Guangyu, Xiaorong Gao, Yijun Wang, Bo Hong, and Shangkai Gao (Nov. 2009). “VEP-based brain-computer interfaces: time, frequency, and code modulations [Research Frontier]”. In: *IEEE Computational Intelligence Magazine* 4.4, pp. 22–26 (cit. on p. 45).

Birbaumer, N, T Elbert, A G Canavan, and B Rockstroh (Jan. 1990). “Slow potentials of the cerebral cortex and behavior.” In: *Physiological Reviews* 70.1, pp. 1–41 (cit. on p. 45).

Birbaumer, N., N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor (Mar. 1999). “A spelling device for the paralysed”. In: *Nature* 398.6725, pp. 297–298 (cit. on p. 46).

Birkholz, Peter, Bernd J. Kroger, and Christiane Neuschaefer-Rube (July 2011). “Model-Based Reproduction of Articulatory Trajectories for Consonant–Vowel Sequences”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.5, pp. 1422–1433 (cit. on p. 82).

Bladon, A and G Fant (1978). “A two-formant model and the cardinal vowels”. In: p. 15 (cit. on p. 16).

Bocquelet, Florent (2017). “Toward a brain-computer interface for speech restoration”. Doctoral dissertation. Université Grenoble Alpes (cit. on pp. 57, 101, 151, 194).

Bocquelet, Florent, Thomas Hueber, Laurent Girin, Pierre Badin, and Blaise Yvert (2014). “Robust articulatory speech synthesis using deep neural networks for BCI applications”. In: *Fifteenth Annual Conference of the International Speech Communication Association* (cit. on pp. 82, 151).

Bocquelet, Florent, Thomas Hueber, Laurent Girin, Stéphan Chabardès, and Blaise Yvert (Nov. 1, 2016a). "Key considerations in designing a speech brain-computer interface". In: *Journal of Physiology-Paris*. SI: GDR Multielectrode 110.4, pp. 392–401 (cit. on pp. 56, 59).

Bocquelet, Florent, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert (Sept. 2016b). *BY2014 articulatory-acoustic dataset*. https://doi.org/10.5281/zenodo.154083 (cit. on pp. 65, 66, 195).

– (Nov. 23, 2016c). "Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces". In: *PLOS Computational Biology* 12.11, e1005119 (cit. on pp. 12, 14, 75, 84, 117).

Boersma, Paul and David Weenink (2021). *Praat: doing phonetics by computer [Computer program]*. Version 6.1.55 (cit. on pp. 77, 197).

Bogert, Bruce P (1963). "The quefrency alanysis of time series for echoes; Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking". In: *Time series analysis*, pp. 209–243 (cit. on p. 22).

Bouchard, Kristofer E., Nima Mesgarani, Keith Johnson, and Edward F. Chang (Mar. 21, 2013). "Functional Organization of Human Sensorimotor Cortex for Speech Articulation". In: *Nature* 495.7441, pp. 327–332 (cit. on pp. 60, 61, 148, 152).

Brumberg, Jonathan S., Kevin M. Pitt, and Jeremy D. Burnison (Apr. 2018). "A Noninvasive Brain-Computer Interface for Real-Time Speech Synthesis: The Importance of Multimodal Feedback". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.4, pp. 874–881 (cit. on p. 59).

BS.1534, ITU recommendation (2014). "Method for the subjective assessment of intermediate quality level of audio systems". In: *R BS.*, p. 36 (cit. on p. 89).

Buzsáki, György, Costas A. Anastassiou, and Christof Koch (June 2012). "The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes". In: *Nature Reviews Neuroscience* 13.6, pp. 407–420 (cit. on p. 148).

Calvetti, D., S. Morigi, L. Reichel, and F. Sgallari (Nov. 1, 2000). "Tikhonov regularization and the L-curve for large discrete ill-posed problems". In: *Journal of Computational and Applied Mathematics*. Numerical Analysis 2000. Vol. III: Linear Algebra 123.1, pp. 423–446 (cit. on p. 94).

Camacho, Arturo and John G. Harris (Sept. 2008). "A sawtooth waveform inspired pitch estimator for speech and music". In: *The Journal of the Acoustical Society of America* 124.3, pp. 1638–1652 (cit. on p. 27).

Cao, Beiming, Myungjong Kim, Jun Wang, Jan Santen, Ted Mau, and Jun Wang (Sept. 2, 2018). "Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information". In: (cit. on p. 150).

Castermans, Thierry, Matthieu Duvinage, Guy Cheron, and Thierry Dutoit (2013). "Towards effective non-invasive brain-computer interfaces dedicated to gait rehabilitation systems". In: *Brain sciences* 4.1, pp. 1–48 (cit. on p. 31).

Chao, Zenas, Yasuo Nagasaka, and Naotaka Fujii (2010). “Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey”. In: *Frontiers in Neuroengineering* 3 (cit. on pp. 49, 93).

Chapin, John K., Karen A. Moxon, Ronald S. Markowitz, and Miguel A. L. Nicolelis (July 1999). “Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex”. In: *Nature Neuroscience* 2.7, pp. 664–670 (cit. on p. 49).

Chartier, Josh, Gopala K. Anumanchipalli, Keith Johnson, and Edward F. Chang (June 2018). “Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex”. In: *Neuron* 98.5, 1042–1054.e4 (cit. on pp. 59, 151).

Chen, Xiaogang, Zhikai Chen, Shangkai Gao, and Xiaorong Gao (Oct. 2, 2014). “A high-ITR SSVEP-based BCI speller”. In: *Brain-Computer Interfaces* 1.3, pp. 181–191 (cit. on p. 45).

Chiba, T and M Kajiyama (1941). “The vowel, its nature and structure Tokyo-Kaiseikan Pub”. In: *Co., Tokyo* (cit. on pp. 16, 19, 20).

Chollet, François et al. (2015). *keras*. `https://keras.io` (cit. on p. 118).

Chu, Wai C (2003). “Foundation and Evolution of Standardized Coders”. In: p. 578 (cit. on p. 21).

Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio (Dec. 11, 2014). “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling”. In: *arXiv:1412.3555 [cs]*. arXiv: `1412.3555` (cit. on p. 86).

Cohen, David (Aug. 23, 1968). “Magnetoencephalography: Evidence of Magnetic Fields Produced by Alpha-Rhythm Currents”. In: *Science* 161.3843, pp. 784–786 (cit. on p. 31).

Collinger, Jennifer L, Brian Wodlinger, John E Downey, Wei Wang, Elizabeth C Tyler-Kabara, Douglas J Weber, Angus JC McMorland, Meel Velliste, Michael L Boninger, and Andrew B Schwartz (Feb. 16, 2013). “High-performance neuroprosthetic control by an individual with tetraplegia”. In: *The Lancet* 381.9866, pp. 557–564 (cit. on pp. 49, 50).

Conant, David F., Kristofer E. Bouchard, Matthew K. Leonard, and Edward F. Chang (Mar. 21, 2018). “Human Sensorimotor Cortex Control of Directly Measured Vocal Tract Movements during Vowel Production”. In: *The Journal of Neuroscience* 38.12, pp. 2955–2966 (cit. on pp. 59, 61).

Cossu, Massimo, Francesco Cardinale, Laura Castana, Alberto Citterio, Stefano Francione, Laura Tassi, Alim L. Benabid, and Giorgio Lo Russo (Oct. 1, 2005). “Stereoelectroencephalography in the Presurgical Evaluation of Focal Epilepsy: A Retrospective Analysis of 215 Procedures”. In: *Neurosurgery* 57.4, pp. 706–718 (cit. on p. 32).

Crone, N. (Dec. 1, 1998). “Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band”. In: *Brain* 121.12, pp. 2301–2315 (cit. on p. 32).

Crone, Nathan E, Dana Boatman, Barry Gordon, and Lei Hao (Apr. 2001). “Induced electrocorticographic gamma activity during auditory perception”. In: *Clinical Neurophysiology* 112.4, pp. 565–582 (cit. on p. 61).

Cunnington, Ross, Robert Iansek, John L. Bradshaw, and Jim G. Phillips (Oct. 1, 1996). "Movement-related potentials associated with movement preparation and motor imagery". In: *Experimental Brain Research* 111.3, pp. 429–436 (cit. on p. 45).

Donchin, E. (1981). "Surprise!... Surprise?" In: *Psychophysiology* 18.5, pp. 493–513 (cit. on p. 42).

Eliseyev, Andrey, Cecile Moro, Jean Faber, Alexander Wyss, Napoleon Torres, Corinne Mestais, Alim Louis Benabid, and Tetiana Aksenova (July 2012). "L1-Penalized N-way PLS for subset of electrodes selection in BCI experiments". In: *Journal of Neural Engineering* 9.4, p. 045010 (cit. on p. 49).

Eriksson, Anders (1995). "The frequency range of the voice fundamental in the speech of male and female adults". In: p. 11 (cit. on p. 16).

Fant, G (1981). "The source filter concept in voice production". In: p. 19 (cit. on pp. 19, 20).

Farwell, L.A. and E. Donchin (Dec. 1988). "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials". In: *Electroencephalography and Clinical Neurophysiology* 70.6, pp. 510–523 (cit. on p. 42).

Fetz, E. E. and D. V. Finocchio (Oct. 22, 1971). "Operant Conditioning of Specific Patterns of Neural and Muscular Activity". In: *Science* 174.4007, pp. 431–435 (cit. on p. 48).

Fetz, Eberhard E. (Feb. 28, 1969). "Operant Conditioning of Cortical Unit Activity". In: *Science* 163.3870, pp. 955–958 (cit. on p. 48).

Flesher, Sharlene N., Jennifer L. Collinger, Stephen T. Foldes, Jeffrey M. Weiss, John E. Downey, Elizabeth C. Tyler-Kabara, Sliman J. Bensmaia, Andrew B. Schwartz, Michael L. Boninger, and Robert A. Gaunt (Oct. 19, 2016). "Intracortical microstimulation of human somatosensory cortex". In: *Science Translational Medicine* 8.361 (cit. on p. 50).

Flesher, Sharlene N., John E. Downey, Jeffrey M. Weiss, Christopher L. Hughes, Angelica J. Herrera, Elizabeth C. Tyler-Kabara, Michael L. Boninger, Jennifer L. Collinger, and Robert A. Gaunt (May 21, 2021). "A brain-computer interface that evokes tactile sensations improves robotic arm control". In: *Science* 372.6544, pp. 831–836 (cit. on p. 50).

Fourier, Jean Baptiste Joseph baron (1822). *Théorie analytique de la chaleur*. Chez Firmin Didot, père et fils (cit. on p. 189).

Fraser, George W, Steven M Chase, Andrew Whitford, and Andrew B Schwartz (Oct. 1, 2009). "Control of a brain–computer interface without spike sorting". In: *Journal of Neural Engineering* 6.5, p. 055004 (cit. on pp. 48, 49).

Fujimura, Osamu (1962). "Analysis of Nasal Consonants". In: p. 12 (cit. on p. 19).

Fukada, T., K. Tokuda, T. Kobayashi, and S. Imai (1992). "An adaptive algorithm for mel-cepstral analysis of speech". In: *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. San Francisco, CA, USA: IEEE, 137–140 vol.1 (cit. on pp. 26, 197).

Georgopoulos, A. P., R. E. Kettner, and A. B. Schwartz (Aug. 1, 1988). “Primate motor cortex and free arm movements to visual targets in three- dimensional space. II. Coding of the direction of movement by a neuronal population”. In: *Journal of Neuroscience* 8.8, pp. 2928–2937 (cit. on p. 48).

Gevins, Alan, Brian Cutillo, John Desmond, Michael Ward, Steven Bressler, Nicholas Barbero, and Kenneth Laxer (July 1, 1994). “Subdural grid recordings of distributed neocortical networks involved with somatosensory discrimination”. In: *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 92.4, pp. 282–290 (cit. on p. 32).

Goncharova, I. I, D. J McFarland, T. M Vaughan, and J. R Wolpaw (Sept. 1, 2003). “EMG contamination of EEG: spectral and topographical characteristics”. In: *Clinical Neurophysiology* 114.9, pp. 1580–1593 (cit. on p. 31).

Gordon, Matthew (Aug. 25, 1998). “The Phonetics and Phonology of Non-modal Vowels: A Cross-Linguistic Perspective”. In: *Annual Meeting of the Berkeley Linguistics Society* 24.1, p. 93 (cit. on p. 7).

Gosztolya, Gábor, Ádám Pintér, László Tóth, Tamás Grósz, Alexandra Markó, and Tamás Gábor Csapó (July 2019). “Autoencoder-Based Articulatory-to-Acoustic Mapping for Ultrasound Silent Speech Interfaces”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. 2019 International Joint Conference on Neural Networks (IJCNN). ISSN: 2161-4407, pp. 1–8 (cit. on p. 150).

Graves, Alex and Jürgen Schmidhuber (July 1, 2005). “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural Networks*. IJCNN 2005 18.5, pp. 602–610 (cit. on p. 85).

Gray, Henry (1878). *Anatomy of the human body*. Vol. 8. Lea & Febiger (cit. on p. 29).

Guenther, Frank H. (2016). *Neural Control of Speech*. The MIT Press (cit. on pp. 35, 36).

Guenther, Frank H., Jonathan S. Brumberg, E. Joseph Wright, Alfonso Nieto-Castanon, Jason A. Tourville, et al. (Dec. 9, 2009). “A Wireless Brain-Machine Interface for Real-Time Speech Synthesis”. In: *PLOS ONE* 4.12, e8218 (cit. on pp. 49, 52, 53, 59, 61).

Guenther, Frank H., Satrajit S. Ghosh, and Jason A. Tourville (Mar. 1, 2006). “Neural modeling and imaging of the cortical interactions underlying syllable production”. In: *Brain and Language* 96.3, pp. 280–301 (cit. on p. 34).

Guenther, Frank H., Michelle Hampson, and Dave Johnson (1998). “A theoretical investigation of reference frames for the planning of speech movements”. In: *Psychological Review* 105.4, pp. 611–633 (cit. on p. 34).

Gussenhoven, Carlos (2004). *The Phonology of Tone and Intonation*, p. 381 (cit. on p. 7).

Han, Wei, Cheong-Fat Chan, Chiu-Sing Choy, and Kong-Pang Pun (May 2006). “An efficient MFCC extraction method in speech recognition”. In: *2006 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2006 IEEE International Symposium on Circuits and Systems (ISCAS). ISSN: 2158-1525, 4 pp.–- (cit. on p. 24).

Haut, Sheryl R., Cynthia Swick, Katherine Freeman, and Susan Spencer (2002). “Seizure Clustering during Epilepsy Monitoring”. In: *Epilepsia* 43.7, pp. 711–715 (cit. on p. 55).

Helmholtz, Hermann L. F. (Oct. 4, 2009). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press. 855 pp. (cit. on p. 19).

Herff, Christian, Dominic Heger, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz (June 12, 2015). "Brain-to-text: decoding spoken phrases from phone representations in the brain". In: *Frontiers in Neuroscience* 9 (cit. on p. 58).

Herff, Christian, Garett Johnson, Lorenz Diener, Jerry Shih, Dean Krusienski, and Tanja Schultz (Aug. 2016). "Towards direct speech synthesis from ECoG: A pilot study". In: IEEE, pp. 1540–1543 (cit. on pp. 58, 61).

Herff, Christian, Dean J. Krusienski, and Pieter Kubben (2020). "The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions". In: *Frontiers in Neuroscience* 14 (cit. on p. 59).

Hickok, Gregory and David Poeppel (May 2004). "Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language". In: *Cognition* 92.1, pp. 67–99 (cit. on p. 33).

– (May 2007). "The cortical organization of speech processing". In: *Nature Reviews Neuroscience* 8.5, pp. 393–402 (cit. on pp. 33, 34, 146, 148).

Hiroya, Sadao and Masaaki Honda (2004). "Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model". In: *IEICE TRANSACTIONS on Information and Systems* 87.5, pp. 1071–1078 (cit. on p. 82).

Hitch, Doug (Jan. 1, 2017). "Vowel spaces and systems". In: *Toronto Working Papers in Linguistics* 38 (cit. on p. 18).

Hochberg, Leigh R., Daniel Bacher, Beata Jarosiewicz, Nicolas Y. Masse, John D. Simeral, et al. (May 2012). "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm". In: *Nature* 485.7398, pp. 372–375 (cit. on pp. 48, 49).

Hochberg, Leigh R., Mijail D. Serruya, Gerhard M. Friehs, Jon A. Mukand, Maryam Saleh, Abraham H. Caplan, Almut Branner, David Chen, Richard D. Penn, and John P. Donoghue (July 2006). "Neuronal ensemble control of prosthetic devices by a human with tetraplegia". In: *Nature* 442.7099, pp. 164–171 (cit. on pp. 48, 50).

Hochreiter and Schmidthuber (1997). "Long short-term memory". In: *Neural Computation* 9(8), pp. 1735–1780 (cit. on p. 85).

Hwang, Han-Jeong, Jeong-Hwan Lim, Young-Jin Jung, Han Choi, Sang Woo Lee, and Chang-Hwan Im (June 30, 2012). "Development of an SSVEP-based BCI spelling system adopting a QWERTY-style LED keyboard". In: *Journal of Neuroscience Methods* 208.1, pp. 59–65 (cit. on p. 45).

Ibayashi, Kenji, Naoto Kunii, Takeshi Matsuo, Yohei Ishishita, Seijiro Shimada, Kensuke Kawai, and Nobuhito Saito (2018). "Decoding Speech With Integrated Hybrid Signals Recorded From the Human Ventral Motor Cortex". In: *Frontiers in Neuroscience* 12 (cit. on pp. 58, 59, 148).

Imai, S (2003). "Speech signal processing toolkit: Sptk version 3.0". In: *http://kt-lab. ics. nitech. ac. jp./˜ tokuda/SPTK/release/SPTKref-3.0. pdf* (cit. on p. 75).

Imai, S. (1983). “Cepstral analysis synthesis on the mel frequency scale”. In: *ICASSP ’83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 8. Boston, MASS, USA: Institute of Electrical and Electronics Engineers, pp. 93–96 (cit. on pp. 25, 26).

Ioffe, Sergey and Christian Szegedy (Feb. 10, 2015). “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *arXiv:1502.03167 [cs]*. arXiv: 1502.03167 (cit. on p. 118).

Jadoul, Yannick, Bill Thompson, and Bart De Boer (Nov. 2018). “Introducing Parselmouth: A Python Interface to Praat”. In: *Journal of Phonetics* 71, pp. 1–15 (cit. on p. 77).

Jarosiewicz, Beata, Anish A. Sarma, Daniel Bacher, Nicolas Y. Masse, John D. Simeral, et al. (Nov. 11, 2015). “Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface”. In: *Science Translational Medicine* 7.313, 313ra179–313ra179 (cit. on pp. 48, 49, 52).

Jin, Zeyu, Adam Finkelstein, Gautham J. Mysore, and Jingwan Lu (Apr. 2018). “Fftnet: A Real-Time Speaker-Dependent Neural Vocoder”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ISSN: 2379-190X, pp. 2251–2255 (cit. on p. 28).

Jobsis, Frans F (1977). “Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters”. In: *Science* 198.4323, pp. 1264–1267 (cit. on p. 30).

Jong, Sijmen de (Mar. 1, 1993). “SIMPLS: An alternative approach to partial least squares regression”. In: *Chemometrics and Intelligent Laboratory Systems* 18.3, pp. 251–263 (cit. on pp. 93, 95).

Jongman, Allard, Ratree Wayland, and Serena Wong (2000). “Acoustic characteristics of English fricatives”. In: *The Journal of the Acoustical Society of America* 108.3, p. 1252 (cit. on p. 17).

Käthner, Ivo, Carolin A. Ruf, Emanuele Pasqualotto, Christoph Braun, Niels Birbaumer, and Sebastian Halder (Feb. 1, 2013). “A portable auditory P300 brain–computer interface with directional cues”. In: *Clinical Neurophysiology* 124.2, pp. 327–338 (cit. on p. 42).

Käthner, Ivo, Selina C. Wriessnegger, Gernot R. Müller-Putz, Andrea Kübler, and Sebastian Halder (Oct. 1, 2014). “Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain–computer interface”. In: *Biological Psychology* 102, pp. 118–129 (cit. on p. 48).

Kellis, Spencer, Kai Miller, Kyle Thomson, Richard Brown, Paul House, and Bradley Greger (Oct. 2010). “Decoding spoken words using local field potentials recorded from the cortical surface”. In: *Journal of neural engineering* 7.5, p. 056007 (cit. on p. 57).

Kello, Christopher T. and David C. Plaut (Oct. 1, 2004). “A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters”. In: *The Journal of the Acoustical Society of America* 116.4, pp. 2354–2364 (cit. on p. 82).

Kiernan, Matthew C, Steve Vucic, Benjamin C Cheah, Martin R Turner, Andrew Eisen, Orla Hardiman, James R Burrell, and Margaret C Zoing (Mar. 12, 2011). “Amyotrophic lateral sclerosis”. In: *The Lancet* 377.9769, pp. 942–955 (cit. on p. 41).

Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (cit. on p. 118).

Kirschfeld, Kuno (Mar. 1, 2005). “The physical basis of alpha waves in the electroencephalogram and the origin of the “Berger effect””. In: *Biological Cybernetics* 92.3, pp. 177–185 (cit. on p. 46).

Klatt, Dennis, John Iles, Nick Ing-Simmonsi, and Reece H. Dunn (2015). *Praat: doing phonetics by computer [Computer program]*. Version 3.05. License: GPL-3.0 (cit. on p. 77).

Klatt, Dennis H. (Mar. 1, 1980). “Software for a cascade/parallel formant synthesizer”. In: *The Journal of the Acoustical Society of America* 67.3, pp. 971–995 (cit. on pp. 77, 197).

Kohnen, R. F., J. C. M. Lavrijsen, J. H. J. Bor, and R. T. C. M. Koopmans (June 1, 2013). “The prevalence and characteristics of patients with classic locked-in syndrome in Dutch nursing homes”. In: *Journal of Neurology* 260.6, pp. 1527–1534 (cit. on p. 40).

Kornhuber, Hans H and Lüder Deecke (1965). “Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale”. In: *Pflüger’s Archiv für die gesamte Physiologie des Menschen und der Tiere* 284.1, pp. 1–17 (cit. on p. 45).

Krauledat, Matthias, Michael Tangermann, Benjamin Blankertz, and Klaus-Robert Müller (Aug. 13, 2008). “Towards Zero Training for Brain-Computer Interfacing”. In: *PLOS ONE* 3.8, e2967 (cit. on p. 47).

Krolak-Salmon, Pierre, Marie-Anne Hénaff, Catherine Tallon-Baudry, Blaise Yvert, Marc Guénot, Alain Vighetto, François Mauguière, and Olivier Bertrand (2003). “Human lateral geniculate nucleus and visual cortex respond to screen flicker”. In: *Annals of Neurology* 53.1, pp. 73–80 (cit. on p. 43).

Kübler, Andrea, Femke Nijboer, Jürgen Mellinger, Theresa M Vaughan, Hannelore Pawelzik, Gerwin Schalk, Dennis J McFarland, Niels Birbaumer, and Jonathan R Wolpaw (2005). “Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface”. In: *Neurology* 64.10, pp. 1775–1777 (cit. on p. 47).

La Vaque, T. J. (Apr. 1, 1999). “The History of EEG Hans Berger”. In: *Journal of Neurotherapy* 3.2, pp. 1–9 (cit. on p. 46).

Ladefoged, Peter and Keith Johnson (2015). *A Course in Phonetics*, p. 355 (cit. on p. 9).

Lee, Bernard S. (Nov. 1, 1950). “Effects of Delayed Speech Feedback”. In: *The Journal of the Acoustical Society of America* 22.6, pp. 824–826 (cit. on p. 122).

Leuthardt, Eric C, Gerwin Schalk, Jonathan R Wolpaw, Jeffrey G Ojemann, and Daniel W Moran (June 1, 2004). “A brain–computer interface using electrocorticographic signals in humans”. In: *Journal of Neural Engineering* 1.2, pp. 63–71 (cit. on pp. 59, 61, 148).

Li, Yunghui, Latifa Nabila Harfiya, Kartika Purwandari, and Yue-Der Lin (Sept. 30, 2020). "Real-Time Cuffless Continuous Blood Pressure Estimation Using Deep Learning Model". In: *Sensors* 20 (cit. on p. 86).

Liu, Zheng-Chen, Zhen-Hua Ling, and Li-Rong Dai (Sept. 8, 2016). "Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks". In: pp. 1502–1506 (cit. on pp. 14, 82, 150).

Livezey, Jesse A., Kristofer E. Bouchard, and Edward F. Chang (Mar. 26, 2018). "Deep learning as a tool for neural data analysis: speech classification and cross-frequency coupling in human sensorimotor cortex". In: *arXiv:1803.09807 [cs, q-bio]*. arXiv: 1803.09807 (cit. on p. 61).

Logothetis, Nikos, Jon Pauls, M.A. Augath, T Trinath, and A Oeltermann (Aug. 1, 2001). "Neurophysiological Investigation of the Basis of the fMRI Signal". In: *Nature* 412, pp. 150–7 (cit. on p. 30).

Lotte, F., M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi (2007). "A review of classification algorithms for EEG-based brain–computer interfaces". In: *Journal of Neural Engineering* 4.2, R1 (cit. on p. 61).

Lotte, Fabien, Jonathan S. Brumberg, Peter Brunner, Aysegul Gunduz, Anthony L. Ritaccio, Cuntai Guan, and Gerwin Schalk (2015). "Electrocorticographic representations of segmental features in continuous speech". In: *Frontiers in Human Neuroscience* 9 (cit. on p. 61).

Lühmann, Alexander von, Christian Herff, Dominic Heger, and Tanja Schultz (2015). "Toward a wireless open source instrument: functional near-infrared spectroscopy in mobile neuroergonomics and BCI applications". In: *Frontiers in human neuroscience* 9, p. 617 (cit. on p. 31).

Maekawa, Kikuo (2002). "From articulatory phonetics to the physics of speech: Contribution of Chiba and Kajiyama." In: *Acoustical Science and Technology* 23.4, pp. 185–188 (cit. on p. 19).

Martin, Stephanie, Peter Brunner, Iñaki Iturrate, José del R. Millán, Gerwin Schalk, Robert T. Knight, and Brian N. Pasley (May 11, 2016). "Word pair classification during imagined speech using direct brain recordings". In: *Scientific Reports* 6, p. 25803 (cit. on pp. 56, 61).

Martin, Stephanie, Iñaki Iturrate, José del R. Millán, Robert T. Knight, and Brian N. Pasley (2018). "Decoding Inner Speech Using Electrocorticography: Progress and Challenges Toward a Speech Prosthesis". In: *Frontiers in Neuroscience* 12 (cit. on p. 56).

Martin, Stéphanie, Peter Brunner, Chris Holdgraf, Hans-Jochen Heinze, Nathan E. Crone, Jochem Rieger, Gerwin Schalk, Robert T. Knight, and Brian N. Pasley (May 27, 2014). "Decoding spectrotemporal features of overt and covert speech from the human cortex". In: *Frontiers in Neuroengineering* 7 (cit. on p. 56).

Maynard, Edwin M., Craig T. Nordhausen, and Richard A. Normann (Mar. 1, 1997). "The Utah Intracortical Electrode Array: A recording structure for potential brain-computer interfaces". In: *Electroencephalography and Clinical Neurophysiology* 102.3, pp. 228–239 (cit. on p. 32).

McFarland, Dennis J, William A Sarnacki, and Jonathan R Wolpaw (June 1, 2010). “Electroencephalographic (EEG) control of three-dimensional movement”. In: *Journal of Neural Engineering* 7.3, p. 036007 (cit. on p. 47).

Mehta, Paul, Wendy Kaye, Leah Bryan, Theodore Larson, Timothy Copeland, Jennifer Wu, Oleg Muravov, and Kevin Horton (2016). “Prevalence of Amyotrophic Lateral Sclerosis — United States, 2012–2013”. In: *Morbidity and Mortality Weekly Report: Surveillance Summaries* 65.8, pp. 1–12 (cit. on p. 41).

Meng, Kevin, David B Grayden, and Mark J Cook (2021). “Identification of discriminative features for decoding overt and imagined speech using stereotactic electroencephalography”. In: p. 6 (cit. on p. 59).

Mensh, B.D., J. Werfel, and H.S. Seung (June 2004). “BCI competition 2003-data set Ia: combining gamma-band power with slow cortical potentials to improve single-trial classification of electroencephalographic signals”. In: *IEEE Transactions on Biomedical Engineering* 51.6, pp. 1052–1056 (cit. on p. 46).

Monti, Martin M., Audrey Vanhaudenhuyse, Martin R. Coleman, Melanie Boly, John D. Pickard, Luaba Tshibanda, Adrian M. Owen, and Steven Laureys (Feb. 18, 2010). “Willful Modulation of Brain Activity in Disorders of Consciousness”. In: *New England Journal of Medicine* 362.7, pp. 579–589 (cit. on p. 41).

Morise, Masanori and Yusuke Watanabe (May 1, 2018). “Sound quality comparison among high-quality vocoders by using re-synthesized speech”. In: *Acoustical Science and Technology* 39.3, pp. 263–265 (cit. on p. 27).

Morise, Masanori, Fumiya YOKOMORI, and Kenji Ozawa (July 1, 2016). “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications”. In: *IEICE Transactions on Information and Systems* E99.D, pp. 1877–1884 (cit. on p. 27).

Moses, David A., Matthew K. Leonard, and Edward F. Chang (Feb. 2018). “Real-time classification of auditory sentences using evoked cortical activity in humans”. In: *Journal of Neural Engineering* 15.3, p. 036005 (cit. on pp. 58, 61, 148).

Moses, David A., Matthew K. Leonard, Joseph G. Makin, and Edward F. Chang (July 30, 2019). “Real-time decoding of question-and-answer speech dialogue using human cortical activity”. In: *Nature Communications* 10.1, p. 3096 (cit. on p. 58).

Moses, David A., Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, et al. (July 15, 2021). “Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria”. In: *New England Journal of Medicine* 385.3, pp. 217–227 (cit. on pp. 49, 54, 58).

Moulin-Frier, Clément, Julien Diard, Jean-Luc Schwartz, and Pierre Bessière (Nov. 1, 2015). “COSMO (“Communicating about Objects using Sensory–Motor Operations”): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems”. In: *Journal of Phonetics*. On the cognitive nature of speech sound systems 53, pp. 5–41 (cit. on pp. 37, 38).

Mugler, Emily M., James L. Patton, Robert D. Flint, Zachary A. Wright, Stephan U. Schuele, Joshua Rosenow, Jerry J. Shih, Dean J. Krusienski, and Marc W. Slutzky (2014). "Direct classification of all American English phonemes using signals from functional speech motor cortex". In: *Journal of Neural Engineering* 11.3, p. 035015 (cit. on pp. 58, 61).

Müller, Johannes (1839). *Über die Compensation der physischen Kräfte am menschlichen Stimmorgan: mit Bemerkungen über die Stimme der Säugethiere, Vögel und Amphibien: Fortsetzung und Supplement der Untersuchungen über die Physiologie der Stimme*. Hirschwald (cit. on p. 19).

Niebergall, Aaron, Shuo Zhang, Esther Kunay, Götz Keydana, Michael Job, Martin Uecker, and Jens Frahm (2013). "Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction". In: *Magnetic Resonance in Medicine* 69.2, pp. 477–485 (cit. on p. 11).

Noll, A. Michael (Feb. 1967). "Cepstrum Pitch Determination". In: *The Journal of the Acoustical Society of America* 41.2, pp. 293–309 (cit. on p. 22).

Nunez, Paul L, Ramesh Srinivasan, et al. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA (cit. on p. 31).

Ogawa, Seiji, Tso-Ming Lee, Asha S Nayak, and Paul Glynn (1990). "Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields". In: *Magnetic resonance in medicine* 14.1, pp. 68–78 (cit. on p. 30).

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (Sept. 19, 2016). "WaveNet: A Generative Model for Raw Audio". In: *arXiv:1609.03499 [cs]*. arXiv: 1609.03499 (cit. on p. 28).

Pandarinath, Chethan, Paul Nuyujukian, Christine H Blabe, Brittany L Sorice, Jad Saab, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson (Feb. 21, 2017). "High performance communication by people with paralysis using an intracortical brain-computer interface". In: *eLife* 6. Ed. by Sabine Kastner, e18554 (cit. on p. 52).

Pandarinath, Chethan, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, et al. (Oct. 2018). "Inferring single-trial neural population dynamics using sequential auto-encoders". In: *Nature methods* 15.10, pp. 805–815 (cit. on pp. 49, 52).

Pasley, Brian N., Stephen V. David, Nima Mesgarani, Adeen Flinker, Shihab A. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang (Jan. 31, 2012). "Reconstructing Speech from Human Auditory Cortex". In: *PLOS Biology* 10.1, e1001251 (cit. on pp. 58, 61).

Pei, Xiaomei, Dennis Barbour, Eric C. Leuthardt, and Gerwin Schalk (Aug. 2011). "Decoding Vowels and Consonants in Spoken and Imagined Words Using Electrocorticographic Signals in Humans". In: *Journal of neural engineering* 8.4, p. 046028 (cit. on p. 56).

Perel, Sagi, Patrick T. Sadtler, Emily R. Oby, Stephen I. Ryu, Elizabeth C. Tyler-Kabara, Aaron P. Batista, and Steven M. Chase (Sept. 1, 2015). "Single-unit activity, threshold crossings, and local field potentials in motor cortex differentially encode reach kinematics". In: *Journal of Neurophysiology* 114.3, pp. 1500–1512 (cit. on p. 49).

Perrone-Bertolotti, Marcela, Jan Kujala, Juan R. Vidal, Carlos M. Hamame, Tomas Ossandon, Olivier Bertrand, Lorella Minotti, Philippe Kahane, Karim Jerbi, and Jean-Philippe Lachaux (Dec. 5, 2012). "How silent is silent reading? Intracerebral evidence for top-down activation of temporal voice areas during reading". In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 32.49, pp. 17554–17562 (cit. on p. 56).

Pfurtscheller, G and A Aranibar (June 1, 1977). "Event-related cortical desynchronization detected by power measurements of scalp EEG". In: *Electroencephalography and Clinical Neurophysiology* 42.6, pp. 817–826 (cit. on p. 46).

Pfurtscheller, G, C Neuper, and G Krausz (Oct. 1, 2000). "Functional dissociation of lower and upper frequency mu rhythms in relation to voluntary limb movement". In: *Clinical Neurophysiology* 111.10, pp. 1873–1879 (cit. on p. 46).

Pfurtscheller, G. (July 1, 1992). "Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest". In: *Electroencephalography and Clinical Neurophysiology* 83.1, pp. 62–69 (cit. on p. 46).

Pfurtscheller, Gert, Brendan Allison, Günther Bauernfeind, Clemens Brunner, Teodoro Solis Escalante, Reinhold Scherer, Thorsten Zander, Gernot Mueller-Putz, Christa Neuper, and Niels Birbaumer (2010). "The hybrid BCI". In: *Frontiers in Neuroscience* 4, p. 3 (cit. on p. 46).

Pfurtscheller, Gert, Robert Leeb, Claudia Keinrath, Doron Friedman, Christa Neuper, Christoph Guger, and Mel Slater (Feb. 3, 2006). "Walking from thought". In: *Brain Research* 1071.1, pp. 145–152 (cit. on p. 47).

Pineda, Jaime A. (Dec. 1, 2005). "The functional significance of mu rhythms: Translating "seeing" and "hearing" into "doing"". In: *Brain Research Reviews* 50.1, pp. 57–68 (cit. on p. 46).

Polich, John (Oct. 1, 2007). "Updating P300: An integrative theory of P3a and P3b". In: *Clinical Neurophysiology* 118.10, pp. 2128–2148 (cit. on p. 42).

REAPER (Sept. 2, 2021). *REAPER: Robust Epoch And Pitch EstimatoR*. original-date: 2014-12-22T23:30:40Z (cit. on p. 27).

Richmond, Korin (2006). "A Trajectory Mixture Density Network for the Acoustic-Articulatory Inversion Mapping". In: *INTERSPEECH*, p. 4 (cit. on p. 82).

Roussel, Philémon, Florent Bocquelet, and Blaise Yvert (Jan. 13, 2021). *Matlab package to assess acoustic contamination of neural electrophysiological data* (cit. on p. 73).

Roussel, Philémon, Gaël Le Godais, Florent Bocquelet, Marie Palma, Jiang Hongjie, et al. (Oct. 2020). "Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception". In: 17.5, p. 056028 (cit. on pp. 68, 74, 107, 108, 196).

Sakoe, Hiroaki and Seibi Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition". In: (cit. on p. 78).

Schimpf, Paul and Hesheng Liu (Feb. 12, 2013). "Localizing sources of the P300 using ICA, SSLOFO, and latency mapping". In: (cit. on p. 43).

Schirrmeister, Robin Tibor, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball (2017). "Short title: Convolutional neural networks in EEG analysis". In: *Arxiv*, p. 58 (cit. on p. 61).

Schoeffler, Michael, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre (Feb. 5, 2018). "webMUSHRA — A Comprehensive Framework for Web-based Listening Tests". In: *Journal of Open Research Software* 6.1, p. 8 (cit. on p. 90).

Schönle, Paul W., Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad (May 1, 1987). "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract". In: *Brain and Language* 31.1, pp. 26–35 (cit. on p. 12).

Schultz, Tanja, Michael Wand, Thomas Hueber, Dean J. Krusienski, Christian Herff, and Jonathan S. Brumberg (Dec. 2017). "Biosignal-Based Spoken Communication: A Survey". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12, pp. 2257–2271 (cit. on p. 29).

Schwemmer, Michael A., Nicholas D. Skomrock, Per B. Sederberg, Jordyn E. Ting, Gaurav Sharma, Marcia A. Bockbrader, and David A. Friedenberg (Sept. 24, 2018). "Meeting brain–computer interface user performance expectations using a deep neural network decoding framework". In: *Nature Medicine* (cit. on p. 61).

Sellers, Eric W., David B. Ryan, and Christopher K. Hauser (Oct. 8, 2014). "Noninvasive brain-computer interface enables communication after brainstem stroke". In: *Science translational medicine* 6.257, 257re7 (cit. on p. 42).

Serruya, Mijail D., Nicholas G. Hatsopoulos, Liam Paninski, Matthew R. Fellows, and John P. Donoghue (Mar. 2002). "Instant neural control of a movement signal". In: *Nature* 416.6877, pp. 141–142 (cit. on pp. 49, 50).

Smith, Eimear and Mark Delargy (Feb. 19, 2005). "Locked-in syndrome". In: *BMJ : British Medical Journal* 330.7488, pp. 406–409 (cit. on p. 40).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: p. 30 (cit. on p. 118).

Stevens, S S, J Volkmann, and E B Newman (1937). "A Scale for the Measurement of the Psychological Magnitude Pitch". In: p. 7 (cit. on p. 23).

Stuart, Andrew, Joseph Kalinowski, Michael P. Rastatter, and Kerry Lynch (May 1, 2002). "Effect of delayed auditory feedback on normal speakers at two speech rates". In: *The Journal of the Acoustical Society of America* 111.5, pp. 2237–2241 (cit. on p. 122).

Sutton, Samuel, Margery Braren, Joseph Zubin, and E. R. John (Nov. 26, 1965). "Evoked-Potential Correlates of Stimulus Uncertainty". In: *Science* 150.3700, pp. 1187–1188 (cit. on p. 42).

Sutton, Samuel, Patricia Tueting, Joseph Zubin, and E. R. John (1967). "Information Delivery and the Sensory Evoked Potential". In: *Science* 155.3768, pp. 1436–1439 (cit. on p. 42).

Taguchi, Fumiaki and Tokihiko Kaburagi (Sept. 2, 2018). “Articulatory-to-speech Conversion Using Bi-directional Long Short-term Memory”. In: *Interspeech 2018*. Interspeech 2018. ISCA, pp. 2499–2503 (cit. on pp. 82, 150).

Takayuki, Arai (2001). “The Replication of Chiba and Kajiyama’s Mechanical Models of the Human Vocal Cavity”. In: *Journal of the Phonetic Society of Japan* 5.2, pp. 31–38 (cit. on pp. 19, 20).

Talkin, David (1995). “A robust algorithm for pitch tracking (RAPT)”. In: (cit. on p. 27).

Tankus, Ariel, Itzhak Fried, and Shy Shoham (Aug. 21, 2012). “Structured neuronal encoding and decoding of human speech features”. In: *Nature Communications* 3.1, pp. 1–5 (cit. on p. 58).

Tian, Qiao, Zewang Zhang, Heng Lu, Ling-Hui Chen, and Shan Liu (Oct. 25, 2020). “FeatherWave: An Efficient High-Fidelity Neural Vocoder with Multi-Band Linear Prediction”. In: *Interspeech 2020*. Interspeech 2020. ISCA, pp. 195–199 (cit. on p. 28).

Toda, Tomoki, Alan W. Black, and Keiichi Tokuda (2008). “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model”. In: *Speech Communication* 50.3, pp. 215–227 (cit. on pp. 14, 82).

Tokuda, Keiichi, Takao Kobayashi, and Satoshi Imai (1994). “Recursive Calculation of Mel-Cepstrum from LP Coefficients”. In: p. 7 (cit. on p. 26).

Tomik, Barbara and Roberto J. Guiloff (Jan. 2010). “Dysarthria in amyotrophic lateral sclerosis: A review”. In: *Amyotrophic Lateral Sclerosis* 11.1, pp. 4–15 (cit. on p. 41).

Tourville, Jason A. and Frank H. Guenther (Aug. 1, 2011). “The DIVA model: A neural theory of speech acquisition and production”. In: *Language and Cognitive Processes* 26.7, pp. 952–981 (cit. on pp. 34, 146, 148).

Townsend, G. and V. Platsko (2016). “Pushing the P300-based brain–computer interface beyond 100 bpm: extending performance guided constraints into the temporal domain”. In: *Journal of Neural Engineering* 13.2, p. 026024 (cit. on p. 43).

Traynor, B. J., M. B. Codd, B. Corr, C. Forde, E. Frost, and O. Hardiman (Feb. 1, 1999). “Incidence and prevalence of ALS in Ireland, 1995–1997: A population-based study”. In: *Neurology* 52.3, pp. 504–504 (cit. on p. 41).

Tremain, Thomas E (1982). “The government standard linear predictive coding algorithm: LPC-10”. In: *Speech Technology*, pp. 40–49 (cit. on p. 21).

Tsoneva, Tsvetomira, Gary Garcia-Molina, and Peter Desain (Mar. 2, 2021). “SSVEP phase synchronies and propagation during repetitive visual stimulation at high frequencies”. In: *Scientific Reports* 11.1, p. 4975 (cit. on p. 43).

Umesh, S., L. Cohen, and D. Nelson (1999). “Fitting the Mel scale”. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). Phoenix, AZ, USA: IEEE, 217–220 vol.1 (cit. on p. 23).

Valin, Jean-Marc and Jan Skoglund (Sept. 15, 2019). “A Real-Time Wideband Neural Vocoder at 1.6kb/s Using LPCNet”. In: *Interspeech 2019*. Interspeech 2019. ISCA, pp. 3406–3410 (cit. on p. 28).

Vidal, J J (June 1973). “Toward Direct Brain-Computer Communication”. In: *Annual Review of Biophysics and Bioengineering* 2.1, pp. 157–180 (cit. on p. 39).

Walter, W Grey, Ray Cooper, VJ Aldridge, WC McCallum, and AL Winter (1964). “Contingent negative variation: an electric sign of sensori-motor association and expectancy in the human brain”. In: *nature* 203.4943, pp. 380–384 (cit. on p. 45).

Wang, Yijun, Ruiping Wang, Xiaorong Gao, Bo Hong, and Shangkai Gao (June 2006). “A practical VEP-based brain-computer interface”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14.2, pp. 234–240 (cit. on p. 43).

Watson, Catherine, William Thorpe, and Xiao Lu (Apr. 1, 2009). “A comparison of two techniques that measure vocal tract shape”. In: *Acoustics Australia* 37, pp. 7–11 (cit. on p. 11).

Wessberg, Johan, Christopher R. Stambaugh, Jerald D. Kralik, Pamela D. Beck, Mark Laubach, John K. Chapin, Jung Kim, S. James Biggs, Mandayam A. Srinivasan, and Miguel A. L. Nicolelis (Nov. 2000). “Real-time prediction of hand trajectory by ensembles of cortical neurons in primates”. In: *Nature* 408.6810, pp. 361–365 (cit. on p. 49).

Willett, Francis R., Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy (May 2021). “High-performance brain-to-text communication via handwriting”. In: *Nature* 593.7858, pp. 249–254 (cit. on pp. 52, 53).

Wodlinger, B, J E Downey, E C Tyler-Kabara, A B Schwartz, M L Boninger, and J L Collinger (Feb. 1, 2015). “Ten-dimensional anthropomorphic arm control in a human brain-machine interface: difficulties, solutions, and limitations”. In: *Journal of Neural Engineering* 12.1, p. 016011 (cit. on pp. 49, 50).

Wolpaw, Jonathan R, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan (June 1, 2002). “Brain–computer interfaces for communication and control”. In: *Clinical Neurophysiology* 113.6, pp. 767–791 (cit. on pp. 32, 39).

Wolpaw, Jonathan R., Dennis J. McFarland, Gregory W. Neat, and Catherine A. Forneris (Mar. 1, 1991). “An EEG-based brain-computer interface for cursor control”. In: *Electroencephalography and Clinical Neurophysiology* 78.3, pp. 252–259 (cit. on p. 47).

Wolpaw, Jonathan R. and Elizabeth Winter Wolpaw, eds. (2012). *Brain-computer interfaces: principles and practice*. Oxford ; New York: Oxford University Press. 400 pp. (cit. on p. 44).

Wrench, Alan (1999). *The MOCHA-TIMIT articulatory database* (cit. on pp. 65, 196).

Zheng, Fang, Guoliang Zhang, and Zhanjiang Song (Nov. 2001). “Comparison of different implementations of MFCC”. In: *Journal of Computer Science and Technology* 16.6, pp. 582–589 (cit. on p. 24).

Zierdt, Andreas, Philip Hoole, and Hans G Tillmann (1999). “DEVELOPMENT OF A SYSTEM FOR THREE-DIMENSIONAL FLESHPOINT MEASUREMENT OF SPEECH MOVEMENTS”. In: p. 6 (cit. on p. 12).

Webpages

Becker, Jason (2021). *Jason Becker: Not Dead Yet [Film Trailer]*. URL: https://youtu.be/wGFDWTC8B8g?t=103 (visited on Sept. 9, 2021) (cit. on p. 41).

International Phonetic Association (2018). Creative Commons Attribution-Sharealike 3.0 Unported License. URL: http://www.internationalphoneticassociation.org/content/ipa-chart (visited on Aug. 21, 2021) (cit. on p. 18).

MadBeppo.com (2021). *The organs of speech (from the neck up)*. URL: https://www.madbeppo.com/french-language/the-organs-of-speech-from-the-neck-up/ (visited on Sept. 23, 2021) (cit. on p. 4).

Walinga, Jennifer and Charles Stangor (2014). *Introduction to Psychology - 1st Canadian Edition*. URL: https://opentextbc.ca/introductiontopsychology/chapter/3-1-the-neuron-is-the-building-block-of-the-nervous-system/ (visited on Sept. 26, 2021) (cit. on p. 30).

List of Figures

Appendix A

A.1 Speech representations

A.1.1 Fourier transforms

A.1.1.1. Spectrum

Any sound signal can be characterized by its frequency content. In a speech sound, when the vocal folds vibrate, they oscillate with a distinct frequency F0. The number of time per second the vocal folds repeat their pattern is their fundamental frequency. If their oscillation was perfectly sinusoidal, they would be characterized by the frequency of this sinusoid, however no oscillation of the vocal fold will ever be exactly the same and moreover they are sharper than a smooth sinusoid. That creates respectively noise and harmonic components. Noise is a random combination of sinusoids and harmonics are the multiples of the fundamental frequency. We've also seen that speech also contains noisy signals, those would not have any defined frequency, only random frequencies with a characteristic spectral repartition of the frequencies: a high pitch noise will contain more high frequencies *ie* higher amplitudes for high frequency components.

Fourier introduced the concept of decomposing a function into a sum of simple trigonometric functions (Fourier, 1822). It states that given an integrable function $x : \mathbb{R} \to \mathbb{C}$, its **Fourier transform** $X : \mathbb{R} \to \mathbb{C}$ is defined by:

$$X(f) \triangleq \int_{-\infty}^{\infty} x(t) e^{-2\pi j t} dt \tag{A.1}$$

x is a function of time and X is a function of frequency, associating to each frequency an amplitude and phase (offset of the sinusoid). As x is a sum of sinusoidal components with amplitudes and phases described by X, it can be reconstructed from X as stipulated by Fourier inversion theorem:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(f) e^{2\pi j f} df \tag{A.2}$$

As an ease of notation we will note $\mathcal{F}\{x(t)\}$ the Fourier transform of $x : \mathbb{R} \to \mathbb{C}$ and $\mathcal{F}^{-1}\{X(f)\}$ the inverse Fourier transform of $X : \mathbb{R} \to \mathbb{C}$.

In computers, sound signals are discretized to be represented as a sequence of numbers. Fortunately, the Fourier transform also work on discrete time signals. In practical applications, the **Discrete Fourier Transform** (**DFT**) approximates the Fourier transform for finite discrete time signal. Given a discrete time signal x of length N uniformly sampled at a period T, the DFT and inverse DFT are computed as follow:

$$X(k) \triangleq \sum_{n=0}^{N-1} x(n) e^{-2\pi j \frac{kn}{N}}, \qquad k = 0, 1, 2, \dots N-1 \tag{A.3}$$

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{2\pi j \frac{kn}{N}}, \qquad n = 0, 1, 2, \dots N-1 \tag{A.4}$$

Where $X(k)$ describes the amplitude and phase of the k^{th} frequency bin $f_k = \frac{2\pi k}{NT}$.

The DFT is rarely implemented using its definition equations as it is a very inefficient way of computing the spectrum of a signal. Assuming special cases of computation (easy to satisfy in practice) it is possible to use instead a class of highly optimized algorithms called **FFT** for **Fast Fourier Transform**. In the manuscript, we used interchangeably DFT, Fourier transform and FFT to refer to the FFT.

A.1.1.2. Spectrogram

The Fourier analysis of a signal provides a powerful way to study its frequency components (or spectrum). The downside is that the analysis does not provide any time information about the signal. In order to analyze the evolution of a signal's spectrum over time, the signal is split into short overlapping **frames**. Each frames of the signal is multiplied by a window function with two properties: 1. falling to 0 on each sides 2. Summing two consecutive windowed frames restores the original non windowed signal. The spectrogram of the signal is obtained by concatenating the FFT of each windowed samples.

We can see from the equation A.3 of the DFT that the number of frequency bins of the transformed signal is equal to the number of samples from the time signal. However in the case of a real signal x and its DFT X, we can show that: 1. $X(k) = \overline{X(N-k)}$ and 2. $f_k = \frac{2\pi k}{N} = -f_{N-k}\ [2\pi]$. That means that DFT of a real signal is made of corresponding positive and negative frequencies of same power $|X(k)|^2 = \left|\overline{X(N-k)}\right|^2$. By limiting the power spectrogram of a real signal of size N to its positive frequencies, we can see that the actual number of frequency bins is $N/2$. This result implies that there is a time/frequency resolution trade-off when computing the DFT. Using shorter frames of signal will improve the time resolution but reduce the number of frequency bins and therefore the frequency resolution of the spectrogram. On the contrary, using larger frames improves the frequency resolution but reduces the time resolution of the spectrogram. It is possible to interpolate the DFT of a signal by adding zeros to the sides of the windowed frames, this is called **zero-padding**. That artificially increases the number of points used to compute the FFT while not changing the time resolution.

In the case of real signals, the DFT is often replaced by a **Discrete Cosine Transform** or **DCT**. It is a transformation closely related to the DFT, only it offers a more efficient representation of real signals by dropping the imaginary part (Ahmed et al., 1974).

A.2 Perceptive tests

A.2.1 Instructions - *French*

This section transcribes the exact instructions (in French) given to the participants before and during the perceptive test.

A.2.1.1. Introduction

Dans ce test vous aurez à écouter et à évaluer plusieurs échantillons sonores.

Si vous souhaitez participer à l'expérience, vous certifiez remplir les conditions suivantes:

Vous êtes majeur, ou avez l'accord d'un parent ou tuteur. Le français est votre langue maternelle. Vous ne présentez ni troubles du langage ni de l'audition. Vous êtes

dans de bonnes conditions d'écoute. C'est à dire dans un environnement calme et sans distractions, avec un casque de bonne qualité.

Le test est en deux parties et dure environ 25 minutes au total, bien que cela puisse être variable d'une personne à l'autre. Votre progression totale dans le test est indiquée par la barre jaune en haut de la page.

Afin d'analyser les résultats il vous sera demandé votre age et votre sexe à la fin du test. Pour commencer le test, munissez vous d'un casque audio et cliquez sur suivant.

Toutes les données recueillies sont anonymes et sécurisées.

A.2.1.2. Volume test

Participants can tune the general volume with a test sentence before starting the test.

A.2.1.3. Part 1

Dans ce test vous aurez à écouter des échantillons sonores et à évaluer leur qualité par rapport à un échantillon de référence. Vous attribuerez une note entre 0 et 100 à chaque échantillon, 100 étant pour une qualité similaire à la référence. Vous pouvez réécouter autant de fois que vous le souhaitez chaque son.

Dans la première partie de ce test, les échantillons sonores seront des variantes de la même phrase. Au milieu de cette phrase se trouve une séquence Voyelle-Consonne-Voyelle, par exemple: "Tu t'appelles ama, c'est ça?" ou "Tu t'appelles uru, c'est ça?". Vous devrez vous focaliser sur cette séquence Voyelle-Consonne-Voyelle pour évaluer la qualité de chaque échantillon sonore.

A.2.1.4. Part 2

Dans la seconde et dernière partie de ce test, vous aurez à nouveau à évaluer la qualité de chaque échantillon par rapport à la référence sur une échelle de 0 à 100. Contrairement à la première partie du test, vous évaluerez à présent les échantillons sur leur qualité générale.

Résumé de la thèse

B

Introduction

Contexte et motivation de la thèse

En France, environ 600 patients souffriraient du locked-in syndrome (LIS), et 5000 à 7000 patients seraient atteints de sclérose latérale amyotrophique (SLA). Nombre de ces patients conservent la capacité cognitive de parler et de communiquer mais perdent la capacité physique de le faire. Lorsque cela est possible, des protocoles de communication utilisant des restes de mouvements volontaires peuvent être utilisés pour interagir avec les patients et améliorer leur qualité de vie, mais ces méthodes restent lentes et nécessitent souvent une assistance.

Récemment, les interfaces cerveau-ordinateur (BCI, pour brain-computer interface) ont alimenté l'espoir de systèmes de communication plus efficaces et d'une indépendance globalement améliorée, avec la possibilité d'être également utilisées par des patients qui ne conservent aucun contrôle moteur. Les BCI non invasives qui ne nécessitent pas d'intervention chirurgicale se concentrent généralement sur la communication par le contrôle d'ordinateurs et la saisie sur un clavier virtuel. Idéalement, cependant, une BCI de communication devrait permettre la production naturelle de la parole. Une telle BCI parole nécessiterait un contrôle fin de plusieurs degrés de liberté, ce qui n'est pas possible avec les BCI non invasives. La recherche sur les BCI moteurs a montré que les enregistrements invasifs peuvent fournir des signaux neuronaux suffisamment bons pour contrôler des membres robotisés à plusieurs degrés de liberté. Une BCI pourrait donc éventuellement être conçue pour produire une parole naturelle à partir d'électrodes invasives.

Etant donné que l'implantation d'un sujet avec des électrodes invasives entraîne des risques médicaux, une BCI pour la parole naturelle ne peut pas être développée directement avec un sujet. Son potentiel doit au préalable être validé sur des enregistrements de patients implantés pour des raisons médicales. Cette thèse vise à décoder la parole dans un cadre hors ligne à partir d'enregistrements invasifs existants, en se concentrant sur des méthodes compatibles en temps réel qui pourraient être utilisées plus tard dans une BCI vocal en boucle fermée. Les conceptions

possibles pour une BCI de la parole naturelle comprennent soit le contrôle des descripteurs acoustiques de la parole, soit le contrôle d'un ensemble d'articulateurs virtuels. Le second nécessite de construire un synthétiseur de parole articulatoire-acoustique. Ces différentes conceptions possibles sont schématisées en bleu sur la figure B.1. Enfin, les patients ne restent que quelques jours à l'hôpital avec des électrodes implantées. Ainsi, seules quelques sessions d'enregistrements sont possibles, ce qui limite le nombre de phrases. C'est une limitation pour l'entraînement de réseaux neuronaux profonds au décodage de la parole, car ils nécessitent de grandes quantité de données. Nous avons étudié dans quelle mesure les méthodes linéaires, qui nécessitent moins de données, peuvent être utilisées pour décoder avec succès la parole à partir de signaux cérébraux. L'étude de ces défis techniques du décodage de la parole à partir de l'activité corticale permettrait de faire un pas de plus vers le prototypage d'une BCI pour la parole naturelle.

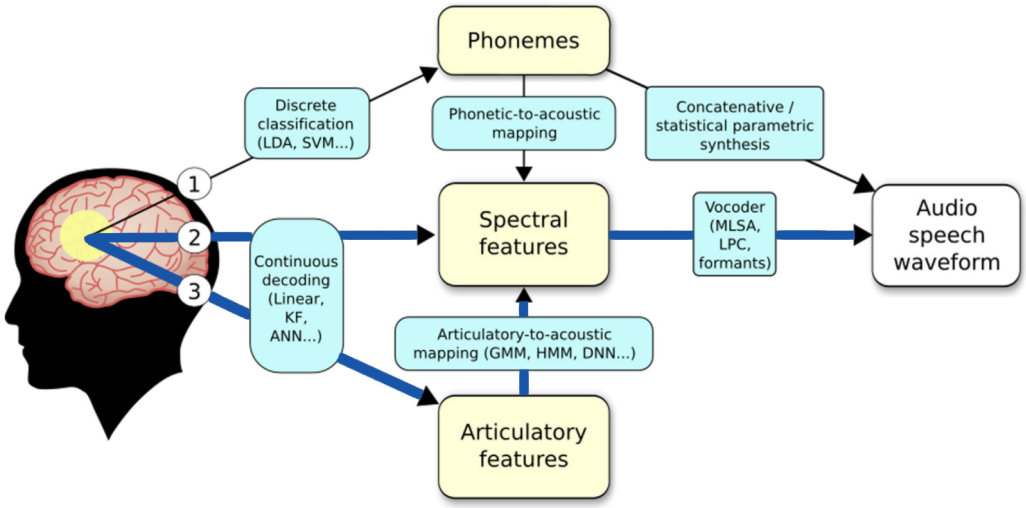


Fig. B.1.: Paradigmes de décodage continus et discrets. Les approches compatibles avec une BCI pour la production de parole naturelle et continue sont marquées par des fléches bleues en gras.
adapté de *(Bocquelet, 2017)*

Objectifs généraux de la thèse

Cette thèse vise à développer des techniques d'apprentissage automatique pour le décodage de la parole à partir d'enregistrements invasifs de l'activité neuronale, en mettant l'accent sur les méthodes linéaires. Elle fait partie d'une tentative à long terme visant à concevoir une BCI qui permettrait à son utilisateur de produire une parole naturelle en temps réel. Le premier objectif de cette thèse est l'étude de méthodes linéaires pour le décodage direct de caractéristiques acoustiques de

la parole à partir de l'activité neuronale, et ce dans un cadre hors ligne. Ceci nécessite de développer un traitement efficace des enregistrements neuronaux et acoustiques existants de la production vocale. En particulier, ce travail examine l'utilisation de représentations spectrales compactes de l'activité corticale, combinée à des techniques automatiques de sélection et de réduction des caractéristiques de l'activité corticale. Les représentations de la parole choisies se concentrent d'une part sur F0 et le mel cepstre pour une synthèse par vocoder, et d'autre part sur F0 et les deux premiers formants pour une synthèse à formants. Ces méthodes doivent toutes être compatibles temps réel afin d'être incluses ultérieurement dans une BCI parole. Le deuxième objectif de cette thèse est de comparer le décodage direct des coefficients acoustiques à partir de l'activité corticale avec un décodage indirect des coefficients acoustiques via une représentation articulatoire. Le décodage indirect nécessite de décoder les trajectoires articulatoires à partir de l'activité corticale, et de prédire la parole acoustique à partir des trajectoires articulatoires décodées. Le décodage des trajectoires articulatoires s'appuie sur les mêmes méthodes linéaires utilisées pour le décodage direct de la parole acoustique, tandis que la synthèse de l'articulation vers l'acoustique se concentre sur des approches basées sur des réseaux de neurones artificiels.

Méthodes

Données

Le travail effectué dans cette thèse a necessité l'utilisation d'enregistrements synchronisés de l'activité corticale, du son produit et de la position des articulateurs lors de la production de parole. Les méthodes d'enregistrement qui ont été utilisées sont l'électrocorticographie (ECoG) pour l'activité cérébrale et l'articulatographie électromagnétique (EMA pour electromagnetic articulography) pour les trajectoires articulatoires. Etant donné que l'enregistrement simultané de signaux ECoG et EMA ne sont pas compatibles dans la pratique, des jeux de données séparés ont été considérés.

Trois jeux de données EMA ont été utilisés. Le principal est BY2014 (Bocquelet et al., 2016b), un large corpus de 676 phrases contenant les enregistrements simultanées de la parole et des trajectoires articulatoires d'un locuteur français natif masculin. Les trajectoires articulatoires ont été enregistrées à 100 Hz en 3 dimensions en 9 points du conduit vocal, les commissures des lèvres; la lèvre supérieure et inférieure; l'avant, le dos et l'arrière de la langue; le voile du palais; et la mandibule. En plus de BY2014, deux autres corpus EMA ont été utilisés: PB2007 et MOCHA-TIMIT

(Wrench, 1999). Le premier contient les enregistrements audio et articulatoires d'un locuteur français natif masculin, et le second d'un locuteur anglais natif masculin.

Trois jeux de données ECoG on été utilisés. Le principal est P5, contenant les enregistrements d'une femme francophone de 38 ans implantée 7 jours pour un bilan épileptiologique pré-chirurgical, la grille ECoG de 72 voies implantée couvrant une large partie de son hémisphère gauche. Durant son séjour, P5 a produit 891 phrases issues de BY2014, dont 391 phrases lues, 250 répétées, et 250 imaginées. En plus de P5, un enregistrement d'un patient francophone masculin de 42 ans (P2) implanté brièvement avec une grille d'ECoG de 256 voies lors d'une chirurgie éveillée a aussi été utilisé. Durant l'enregistrement, P2 a lu 118 phrases de BY2014. Pour finir, un enregistrement d'une patiente anglophone de 30 ans (EC61) implantée pour un bilan épileptiologique à également été utilisé. EC61 a été implanté avec une grille ECoG de 256 voies recouvrant son héisphère droit (dominant), et a lu un total de 460 phrases issues de MOCHA-TIMIT.

Traitement des enregistrements cérébraux

Afin de retirer les artefacts tels que le bruit de ligne, une **référence médiane** a été appliquée aux enregistrements cérébraux. Le spectrogramme de l'activité a été calculé à 100 Hz sur un total de 20 bandes de fréquences de 10 Hz, de 0 à 200 Hz. A cela a été ajouté une charactéristique représentant les potentiels corticaux lents calculée en filtrant l'activité cérébrale entre 0.5 et 5 Hz.

Au cours de sa thèse, Philémon Roussel a montré que certains enregistrements ECoG pouvaient être **contaminés par des signaux acoustiques**. En particulier, certaines éléctrodes peuvent enregistrer le signal acoustique de la parole, ce qui biaiserait toute tentative de décodage de la parole des enregistrements corticaux. Il a été montré par Philémon Roussel que les enregistrements utilisés de P5 et EC61 ne contenaient pas de contamination acoustique. Sur P2 en revanche, de la contamination a été trouvée à partir de 100 Hz ce qui correspond aux valeurs basses de F0 (Roussel et al., 2020). J'ai testé comment une contamination au-delà de 100Hz pouvait influencer le décodage de la parole en retirant les characteristiques spectrales de l'activité cérébrale supérieures à 90 Hz.

Les charactéristiques spectrales de l'activité corticale ne contenant pas d'information sur la parole ont été automatiquement retirées en utilisant un ***t*-test de Welch avec correction Bonferroni**. Pour cela, deux jeux de données contenant des enregistrements cérébraux uniquement lors de phases de parole (production) et hors des phase de parole (ni perception, ni production) ont été comparés. Seules les

characteristiques dont la distribution statistique changeait significativement entre les deux conditions ($p<0.05$) ont été **sélectionnées**.

La production acoustique de la parole et ses processus congnitifs associés ne sont pas synchronisés. En effet, le contrôle motor des articulateurs nécessite une planification en amont, tandis que les processus cognitifs prenant en compte le retour auditif et somatosensoriel ont lieu après la production acoustique. Pour prendre en compte ces phénomènes dans le décodage de la parole, deux méchanismes ont été étudiés: 1. un **délai temporel** variable entre les charactéristiques de l'activité corticale et celles de la parole acoustique/articulatoire, et 2. une **contexte temporel** variable qui consiste à concaténer plusieurs trames consécutives de charactéristiques de l'activité corticale pour décoder une trame de charactéristiques acoustiques/articulatoires de la parole.

Traitement des enregistrements acoustiques

Les signaux acoustiques de la parole ont été décomposés en une représentation source-filtre basée sur F0 et le mel cepstre permettant une synthèse en temps réel par filtre MLSA (Fukada et al., 1992). Le mel cepstre d'ordre 24 et F0 ont été extraits des signaux acoustiques avec SPTK, de sorte à obtenir un signal échantilloné à 100 Hz. Par ailleurs, les deux premiers formants de la parole ont été extraits à 100 Hz par Mohamed Baha Ben Ticha, doctorant au laboratoire, en utilisant la bibliothèque Python Parselmouth qui est une API pour PRAAT (Boersma and Weenink, 2021). Ces deux formants sont suffisants pour décrire les voyelles et, en les combinant à F0, à effectuer une sythèse par Klatt (Klatt, 1980).

Traitement des enregistrements articulatoires

Les enregistrements EMA en 3 dimensions des trajectoires articulatoires de BY2014 contiennent beaucoup de redondances. Ils ont donc été projettés sur le plan mid-sagittal par PCA. Par ailleurs, les commissures des lèvres ont été retirées, ce qui laisse 14 charactéristiques articulatoires à 100 Hz pour BY2014.

Les trajectoires articulatoires de P5 ont été estimées à partir de BY2014 en utilisant un algorithme de déformation temporelle dynamique (DTW, pour dynamic time warping). Les déformations optimales transformant les phrases de BY2014 vers les phrases correspondantes de P5 on été calculées en alignant le mel cepstre, F0, et un booléen charactérisant les phases de paroles et de silences. Ces déformations ont

ensuite été appliquées aux trajectoires articulatoires de BY2014 pour estimer celles de P5.

Synthèse articulatoire

Sous la direction de Thomas Hueber et Laurent Girin au Gipsa-Lab, différentes méthodes de synthèse articulatoire-acoustique on été étudiées, à la fois avec des évaluations objectives et perceptives. Parmi les méthodes étudiées, des réseaux de neurones à propagation avant avec un contexte temporel variable (DNN) et un réseau de neurones bidirectionnel récurrent (biGRU). Les premiers permettent une synthèse compatible temps réel qui ne prend en compte qu'une information limitée dans le temps, tandis que le second synthétise les phrases une par une en prennant une information temporelle indéfiniment longue. Ces méthodes ont été entrainées et évaluées dans une validation croisée par l'erreur quadratique moyenne et la distortion mel cepstrale des synthèses produites. Ensuite, un DNN sans contexte, un DNN avec un contexte optimal (d'après les mesures objectives), et un biGRU ont été comparés perceptivement lors d'un test MUSHRA en ligne incluant également le stimulus original aux différentes synthèses ainsi qu'un synthèse de mauvaise qualité servant d'ancre basse. Le test a été découpé en deux blocs d'environ 25 minutes, l'un évaluant 9 séquences de voyelle-consonne-voyelle et l'autre 10 courtes phrases de BY2014.

Décodage de la parole

Deux méthodes ont été comparées pour réduire le nombre total de charactéristiques spectrales de l'activité corticale, qui peut autrement être très grand, même après sélection. La première est l'**analyse en composantes principales** (PCA), en conservant de 10 à 200 composantes avant calcul du contexte temporel. La seconde est la **réduction par moindre carrés partiels**, en calculant de 3 à 100 composantes après calcul du contexte temporel.

Trois méthodes de regression combinées à une réduction de charactéristiques par PCA ont été comparées, une **simple régression linéaire** et deux régressions linéaires avec **régularisation ridge** en calculant le paramètre de régularisation soit avec une courbe L soit par validation croisée. Ces méthodes ont aussi été comparées à une régression PLS, qui combine une régression linéaire et une régression de charactéristiques par moindre carrés.

Toutes ces méthodes ont été comparées pour le décodage direct des charactéristiques acoustiques et articulatoires de la paroles. Le mel cepstre de P5 a aussi été prédit indirectement des trajectoires articulatoires décodées en utilisant un DNN avec 10 trames de contexte passé et 1 trame de contexte futur. Le DNN a été au préalable entrainé à effectuer une synthèse articulatoire sur BY2014 et éventuellement affiné sur P5. Toutes les méthodes ont été évaluées dans une validation croisée avec 10 blocs.

Résultats

Charactéristiques de l'activité corticale pour le décodage de la parole

La sélection automatique de charactéristiques de l'activité corticale a permis d'en conserver 2687 sur 5376 pour P2, et 1455 sur 1512 pour P5. Pour P2, cette sélection a permis de retirer des électrodes bruitées qui avaient déjà été repérées manuellement, et d'augmenter significativement la corrélation des coefficients mel cepstraux décodés de 0.24 à 0.29. Pour P5, presque toutes les électrodes ont été sélectionnées sur l'ensemble du corpus. En revanche, en séparant les journées d'enregistrement 669, 1114 et 702 charactéristiques ont été sélectionnéese pour les jours 1, 2 et 3, respectivement. Dans tous les cas, le décodage des coefficients mel cepstraux de P5 n'a que peu été affecté par la sélection de charactéristiques.

Pour P5, le décodage des coefficients mel cepstraux, de F0 et des trajectoires articulatoires a été significativement améliorée par l'augmentation du contexte temporel, jusqu'à 210ms la valeur maximale testée. Par ailleurs un délai de 100ms entre les données corticales et acoustiques/articulatoires a été montré optimal, soit une utilisation des 200 dernières ms de l'activité corticale pour décoder la parole à un instant donné.

Une cartographie des charactéristiques sélectionnées ($p<0.05$) de P5 n'a pas donnée d'information clairement interprétable. En revanche, les correlations des coefficients mel cepstraux décodés avec ceux extraits du corpus était significativement meilleures en utilisant uniquement les électrodes temporales qu'en utilisant uniquement les électrodes frontales. De même pour F0. Le décodage des trajectoires articulatoires a au contraire été significativement meilleur en utilisant les électrodes frontales que les électrodes temporales. Ces résultats sont cohérents avec la compréhension actuelle des méchanismes cognitifs de production de la parole. Nous avons également montré que le décodage de la parole était dans les cas significativement meilleur en utilisant l'ensemble des électrodes.

Le décodage des coefficients mel cepstraux de P2 et P5 ont été amélioré par l'utilisation de toutes les bandes de fréquences de l'activité corticale, jusqu'à 200 Hz, par rapport à l'utilisation des bandes de fréquences uniquement sous 90 Hz. Cela est cohérent avec la litérature et avec la cartographie de la sélection des charactéristiques de l'activité corticale de P5. Si l'augmentation était plus importante pour P2 que pour P5, cette analyse ne permet pas d'établir de relation de cause à effet avec la contamination acoustique de P2 dans les hauts gammas (>90 Hz).

Comparison des méthodes de décodage de la parole

Nos résultats ont montré qu'augmenter le nombre de composantes PCA jusqu'à 200 améliorait le décodage, il n'a pas été possible de tester plus de composantes par manque de RAM. Pour la régression PLS, le nombre optimal de composantes a été établi à 12 pour P5, 12 pour le mel cepstre de EC61, et 18 pour F0 de EC61. Toutes les méthodes de linéaire de décodage des charactéristiques acoustiques/articulatoire de la parole ont eu des performances similaires pour P5, ce qui indique d'une part que la régularisation de la régression n'est pas nécessaire de ce cas, et d'autre part que la réduction de charactéristiques par PLS est plus efficace que la réduction par PCA.

L'évaluation objective des différentes méthodes de synthèse acoustique-articulatoire a montré qu'un DNN avec 10 trames de contexte passé et une trame de contexte futur était optimal sur BY2014 et dépassait les performances d'un DNN sans contexte ou d'un biGRU. L'évaluation perceptive n'a pas montré de différences claire dans la qualité de synthèse entre le DNN avec contexte optimal et le biGRU, en revanche ces deux méthodes ont été significativement mieux évaluées que le DNN sans contexte. Le DNN avec contexte optimal a donc été choisi pour le décodage indirect de la parole à partir de l'activité corticale. Cette méthode ayant par ailleurs l'avantage d'être compatible temps réel.

Le décodage indirect du mel cepstre de P5 en utilisant un DNN avec 100 ms de contexte passé et 10 ms de contexte futur a montré une performance similaire au décodage direct pour les régressions linéaire avec PCA (correlations de 0.45), et légèrement moins bonne pour la PLS (correlations de 0.41 pour le décodage indirect, 0.45 pour le décodage direct). Ces performances ont été obtenues après affinage du DNN sur le corpus de trajectoires estimées de P5, sans quoi les corrélations obtenues étaient significativement plus faibles (correlation mediane de 0.07, $p<0.001$).

Le décodage de F0 et des formants avec les méthodes linéaires ne permet pas de correctement prédire les discontinuitées entre phases voisées et non voisées. En

entrainant la regression à prédire uniquement les phases voisées et en le combinant avec un un classifieur parfait du voisement, nos résultats préliminaires montrent une amélioration du décodage. Un véritable classifieur n'a pas été implémenté dans cette thèse. Pour finir, j'ai adapté et testé sur EC61 une méthode de décodage à base de réseaux de neurones artificiels développée par Xingchen Ran (doctorant visiteur au laboratoire). Cette méthode montre une amélioration significative du décodage par rapport à une régression PLS (correlation mediane du mel cepstre: 0.46 et 0.36 respectivement, $p<0.001$).

Perspectives

Le travail présenté dans cette thèse est une étape vers une interface cerveau-ordinateur produisant une parole naturelle. Les méthodes de décodage direct et indirect utilisant des méthodes linéaires étudiées sont toutes compatibles temps réel et pourraient être utilisée pour une BCI en boucle fermée. En particulier la régression PLS a montré une performance compétitive tout en réduisant le nombre de degrés de liberté. Le décodage direct comme indirect ont montré des performances similaires qui ne permettent pas à ce jour de conclure quelle serait la meilleure méthode à implémenter dans une BCI parole. Ces méthodes devraient à présent être implémentées en temps réel pour une future évaluation en boucle fermée. Parmi les pistes futures de travail, un classifieur du voisement devrait être développé, ainsi que des méthodes de décodages basées sur des réseaux de neurones compatibles temps-réel.