Toward a brain-computer interface for speech restoration

Florent Bocquelet

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Biotechnologie, Instrumentation, Signal**

Arrêté ministériel : 25 mai 2016

Présentée par

Florent BOCQUELET

Thèse dirigée par **Blaise YVERT, Directeur de recherche, INSERM**, et
codirigée par **Laurent GIRIN, Professeur des Universités, Université Grenoble Alpes**

préparée au sein des laboratoires **BrainTech** et **Gipsa-lab**
dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

Vers une interface cerveau-machine pour la restauration de la parole

Thèse soutenue publiquement le **24 avril 2017**,
devant le jury composé de :

M. Stéphan CHABARDES
Professeur (CHUG), Membre
M. Frank GUENTHER
Professeur (Université de Boston), Rapporteur
M. Thomas HUEBER
Chargé de recherche (CNRS), Membre
M. Oliver MÜLLER
Professeur (Université de Freiburg), Membre
Mme. Tanja SCHULTZ
Professeur (Université de Brême), Rapportrice
Mme. Agnès TREBUCHON
Maitre de conférences et praticien hospitalier (APHM), Présidente

Abstract

Restoring natural speech in paralyzed and aphasic people could be achieved using a brain-computer interface controlling a speech synthesizer in real-time. The aim of this thesis was thus to develop three main steps toward such proof of concept.

First, a prerequisite was to develop a speech synthesizer producing intelligible speech in real-time with a reasonable number of control parameters. Here we chose to synthesize speech from movements of the speech articulators since recent studies suggested that neural activity from the speech motor cortex contains relevant information to decode speech, and especially articulatory features of speech. We thus developed a speech synthesizer that produced intelligible speech from articulatory data. This was achieved by first recording a large dataset of synchronous articulatory and acoustic data in a single speaker. Then, we used machine learning techniques, especially deep neural networks, to build a model able to convert articulatory data into speech. This synthesizer was built to run in real time. Finally, as a first step toward future brain control of this synthesizer, we tested that it could be controlled in real-time by several speakers to produce intelligible speech from articulatory movements in a closed-loop paradigm.

Second, we investigated the feasibility of decoding speech and articulatory features from neural activity essentially recorded in the speech motor cortex. We built a tool that allowed to localize active cortical speech areas online during awake brain surgery at the Grenoble Hospital and tested this system in two patients with brain cancer. Results show that the motor cortex exhibits specific activity during speech production in the beta and gamma bands, which are also present during speech imagination. The recorded data could be successfully analyzed to decode speech intention, voicing activity and the trajectories of the main articulators of the vocal tract above chance.

Finally, we addressed ethical issues that arise with the development and use of brain-computer interfaces. We considered three levels of ethical questionings, dealing respectively with the animal, the human being, and the human species.

Résumé

Restorer la faculté de parler chez des personnes paralysées et aphasiques pourrait être envisagée via *l'utilisation d'une interface cerveau*-machine permettant de contrôler un synthétiseur de parole en temps réel. *L'objectif de cette thèse était de développer trois aspects nécessaires à la mise au point d'une telle* preuve de concept.

Premièrement, un synthétiseur permettant de produire en temps-réel de la parole intelligible et controlé par un nombre raisonable de paramètres est nécessaire. Nous avons choisi de synthétiser de la parole à partir des mouvements des ar*ticulateurs du conduit vocal. En effet, des études récentes ont suggéré que l'activité neuronale du cortex moteur de la parole pourrait contenir suffisamment d'information pour décoder la parole, et particulièrement ses propriété articulatoire (ex. l'ouver*ture des lèvres). Nous avons donc développé un synthétiseur produisant de la parole intelligible à partir de données articulatoires. Dans un premier temps, nous avons enregistré un large corpus de données articulatoire et acoustiques synchrones chez un locuteur. Ensuite, *nous avons utilisé des techniques d'apprentissage automatique, en particulier des réseaux de neurones profonds,* pour construire un modèle permettant de convertir des données articulatoires en parole. Ce synthétisuer a été construit pour fonctionner en temps réel. Enfin, comme première étape vers un contrôle neuronal de ce *synthétiseur, nous avons testé qu'il pouvait être contrôlé en temps réel par plusieurs locuteurs, pour produire de* la parole inetlligible à partir de leurs mouvements articulatoires dans un paradigme de boucle fermée.

Deuxièmement, nous avons étudié le décodage de la parole et de ses propriétés articulatoires à partir *d'activités neuronales essentiellement enregistrées dans le cortex moteur de la parole. Nous avons construit* un *outil permettant de localiser les aires corticales actives, en ligne pendant des chirurgies éveillées à l'hôpital de Grenoble, et nous avons testé ce système chez deux patients atteints d'un cancer du cerveau. Les résultats ont* montré que le cortex moteur exhibe une activité spécifique pendant la production de parole dans les bandes beta *et gamma du signal, y compris lors de l'imagination de la parole. Les données enregistrées ont ensuite pu être analysées pour décoder l'intention de parler du sujet (ré*elle ou imaginée), ainsi que la vibration des cordes vocales et les trajectoires des articulateurs principaux du conduit vocal significativement au dessus du niveau de la chance.

Enfin, nous nous sommes intéressés aux questions éthiques qui accompagnent le *développement et l'usage* des interfaces cerveau-machine. Nous avons en particulier considéré trois niveaux de réflexion éthique concernant *respectivement l'animal, l'humain et l'humanité.*

Acknowledgement

Content

List of figures

List of tables

Acronyms and terms

1D	One-Dimensional
2D	Two-Dimensional
3D	Tree-Dimensional
ANN	Artificial Neural Network
ALS	Amyotrophic Lateral Sclerosis
ASR	Automatic Speech Recognition
BCI	Brain-Computer Interface
CG	Conjugate Gradient
CNS	Central Nervous System
DAE	Deep Auto-Encoder
DBN	Deep Belief Network
DNN	Deep Neural Network
DoF	Degrees of Freedom
DTW	Dynamic Time Warping
ECoG	ElectroCorticoGraphy
EEG	ElectroEncephaloGraphy
EGG	ElectroGlottoGraph
EMA	ElectroMagnetic Articulography
EMG	ElectroMyoGraphy
EPG	ElectroPalatoGraphy
FMC	Face Motor Cortex
fMRI	functional Magnetic Resonance Imaging
fNIRS	function Near-InfraRed Spectroscopy
GD	Gradient Descent
GMM	Gaussian Mixture Model
GMR	Gaussian Mixture Regression
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
LFP	Local Field Potential
MCD	Mel-Cepstral Distortion
MEA	Micro-Electrodes Array
MEG	MagnetoEncephaloGraphy
MEL	Mel-cepstrum
MLE	Maximum Likelihood Estimation
MLSA	Mel-Log Spectrum Approximation
MMSE	Minimum Mean Squared Error
MRI	Magnetic Resonance Imaging
MRSE	Mean Root Squared Error
MSE	Mean Squared Error
PCA	Principal Component Analysis
Pdf	Probability density function
PET	Positron Emission Tomography
rTMS	repetitive Transcranial Magnetic Stimulation
SD	Standard Deviation
SNR	Signal to Noise Ratio
STFT	Short Term Fourier Transform
STG	Superior Temporal Gyrus
SVM	Support Vector Machine
VCV	Vowel-Consonant-Vowel sequence
vSMC	ventral half of the lateral SensoriMotor Cortex

Phonetic notation

All the phonetic transcriptions contained in this thesis manuscript are using the International Phonetic Alphabet (IPA) (Kenyon, 1929).

Introduction

Motivation of research

*"*Vous n'imaginez pas la gymnastique effectuée machinalement par votre langue pour produire tous les sons du français. Pour l'instant je bute sur le « L », piteux rédacteur en chef *qui ne sait plus articuler le nom de son propre journal"* (*"You cannot imagine the gymnastics* automatically performed by your tongue to produce all the French sounds. For now, I stumble *on the "L", pitiful editor who doesn't know how to articulate the name of his own newspaper."*). This is an extract from the book "Le scaphandre et le papillon", written by Jean-Dominique Bauby using only the blink of his left eye (about 200.000 for the whole book), while suffering from the locked-in syndrome.

In France, about 300.000 people suffer from a strong speech disorder or aphasia that can often occur after a brain stroke but also in case of severe tetraplegia, locked-in syndrome, neurodegenerative diseases such as Amyotrophic Lateral Sclerosis or Parkinson's disease, myopathies, or coma. Some of them are not able to communicate at all while they retain cognition and sensation. For these people, speech loss is an additional affliction that worsens their condition: it makes the communication with caregivers very difficult, and more generally, it can lead to profound social isolation and even depression. Therefore, it is crucial for these patients to restore their ability to communicate with the external world.

Current approaches can provide ways to communicate, mostly through a typing process, by analyzing residual eye movements or brain responses to specific stimuli. However, up to several minutes are needed to type a full sentence, while it only requires about three seconds using natural speech, and not all patients can benefits from these systems.

Indeed, speech remains our most natural and efficient way of communicating. But as Jean-Dominique Bauby mentioned, speech is the result of complex muscle movements, controlled by our nervous system. Over the past decades, Brain-Computer Interfaces (BCI) approaches have been increasingly developed to control the motor movements of effectors (e.g., robotic arms or computer screen cursor) with increasing precision, first in animals and more recently in humans. These systems first enabled controlling a small number of degrees of freedom (DoF), typically 1 or 2, while recent studies have reported that subjects were able to simultaneously control up to 10 DoF in complex motor tasks after appropriate training. However, there has been so far no demonstration of the feasibility to restore direct speech using a BCI approach.

The overall objective of this thesis was thus to develop a parametric speech synthesizer that could be used in a BCI paradigm and to design and run clinical trials in order to collect, analyze and decode speech-related brain activity.

There are several difficulties to overcome in order to reach this goal. First, a speech synthesizer suitable for a BCI approach should run in real-time and be robust enough in order to compensate for brain activity decoding errors. Secondly, speech relies on a high number of DoF for which timing is crucial, which are both challenges for BCIs. Third, many brain areas

are involved in speech production, and recording of the full brain activity with enough time and space resolution is not feasible yet. Fourth, the access to patients for clinical trials is very limited, especially when neural recording requires surgery, which makes it hard to collect large data sets.

With the development of machine learning techniques and data acquisition methods for speech-related signals, parametric speech synthesis is now possible using statistical models of huge data sets. Moreover, recent works showed that brain-computer interfaces now allow control of several DoF, up to around 10, with increasing precision, while using neural activity recorded in small brain areas. These advances are first steps toward a brain-computer interface for speech rehabilitation.

Organization of the manuscript

The thesis manuscript is organized as follows.

Part 1 introduces the topics covered by this thesis and their literature, and is divided into three chapters:

- **Chapter 1** Introduces brain computer interfaces for communication and the key points to consider when developing such brain-computer interface, including the different cortical areas involved during speech production, the different neural activity recording techniques, as well as the different strategies to decode neural activity.
- **Chapter 2** is focused on existing approaches for parametric speech synthesis, and more particularly for articulatory-based parametric speech synthesis. It covers as well common ways of acquiring articulatory data.

Part 2 briefly presents the goal of this thesis.

Part 3 describes the articulatory-based speech synthesizer developed during this thesis:

- **Chapter 3** describes two articulatory-acoustic data sets that were used to perform the speech synthesis. The first corpus was already existing while the second one was specifically recorded from a “reference speaker” for our purpose. Electromagnetic articulography (EMA) was used to acquire articulatory data of the speaker synchronously with the audio speech signal, which was parametrized using mel-cepstrum coefficients.
- **Chapter 4** focuses on speech synthesis from articulatory data from a “reference speaker” and its evaluation. The mapping from articulatory to acoustic data was performed using a Deep Neural Network (DNN) trained on an articulatory-acoustic dataset. This approach was then evaluated using both objective and perceptive listening tests, and compared to a state-of-the-art approached based on Gaussian Mixture Regression (GMR).

- **Chapter 5** describes the adaptation of the articulatory speech synthesizer of the "reference speaker" in order to control it in real-time using the articulatory data from new speakers. In a preliminary study, the robustness of the DNN-based articulatory speech synthesizer was assessed using artificially degraded articulatory data as input for the synthesis, and compared to a state-of-the-art GMM model. The articulatory speech synthesizer was adapted to new speakers in order to be controlled in real-time while they were silently articulating.

Part 4 describes preliminary results on analyzing and decoding speech-specific neural activity:

- **Chapter 6** presents a way to automatically localize speech-specific brain areas during awake surgery. Such localization is needed in order to optimize the positioning of micro-electrode arrays that can only cover a limited surface of the cortex.
- **Chapter 7** presents preliminary results on decoding speech features from neural activity. In particular, this chapter is focused on speech intention detection, which consists in predicting, from the neural activity, when a patient intends to speak or not, as well as on voicing activity detection – predicting if the vocal folds are vibrating or not – and the decoding of articulatory trajectories.

Part 5 is an attempt at analyzing the ethical implications of brain-computer interfaces in general and their development.

Finally, **Part 6** summarizes the contributions of this thesis and discusses suggestions for future work.

Part 1: State of the art

Chapter 1: Brain-computer interfaces for speech rehabilitation

I. Introduction: Brain-computer interfaces for communication

Different solutions for restoring communication in patients with severe paralysis have been developed, most often through a typing process in which letters are selected one by one by exploiting residual physiological signals, such as tracking the eyes direction to control a computer mouse cursor and detecting eye blinks to allow the user to click on the letter pointed by the cursor. However, such solutions are only available for patients with sufficient remaining motor control and only allow to control devices with a small number of degrees of freedom. To overcome this problem, communication systems controlled directly by brain signals have thus started to be developed.

This concept has been pioneered by Farwell and Donchin who proposed a spelling device based on the evoked potential P300 (Farwell and Donchin, 1988), a method that has since been used successfully by a patient with amyotrophic lateral sclerosis (ALS) to communicate (Sellers et al., 2014). The P300 is an event related potential generally elicited when a low-probability expected event occurs during a series of high-probability events and can be recorded using electro-encephalography (EEG). It occurs for instance when a subject actively detects a different sound among a series of identical sounds. Other EEG-based approaches use steady-state potentials tuned at different frequencies (Middendorf et al., 2000). When the retina is exposed to a visual stimulation at a specific frequency (generally from 3 to 75Hz), the brain generates visual evoked potentials at identical frequency called steady state visual potential (SSVP). This natural phenomenum can be exploited by displaying on a screen all the letters blinking at different frequencies so that when the patient focuses on a specific letter, the corresponding SSVP can be detected and the letter identified. These EEG-based approaches present the great advantage of being non-invasive. However, they have been limited by a low spelling speed of a few characters per minute, although recent improvements suggest that higher speed could be achieved (Townsend and Platsko, 2016). Moreover, such tasks are very demanding for the subjects that must remain focused and concentrated during the whole typing process (Käthner et al., 2014; Baykara et al., 2016), thus limiting the use of the device over extensive periods of time.

On the other hand, BCI systems based on intracortical recording, while having the major drawback of being invasive, seem to require less concentration effort from the subject, the external device becoming progressively embodied after a period of training (Hochberg et al., 2006, 2012; Collinger et al., 2013; Wodlinger et al., 2014). Moreover, intracortical recordings allow to capture more information and thus lead to a more precise decoding of the user's intention. Combining intracortical recording with self-recalibrating algorithms was recently

shown to allow typing of about 20-30 characters per minute by people with severe paralysis over long periods of use (Jarosiewicz et al., 2015).

However, these typing BCI systems are generally controlled by neural activity recorded from the hand and/or arm area of the motor cortex and are thus an indirect way of communicating that can exploit non-speech brain activity. Such strategy could prevent to communicate while intending another motor action, such as talking while grabbing an object which could prevent combining multiple BCIs using the same brain areas. Indeed, neural activity from the hand or arm cortex areas would be needed to decode a grab action and could not be used at the same time to communicate. Moreover, speech still remains our natural and most efficient way of communicating. In this thesis, we thus envisioned a "speech BCI" that would restore continous speech by decoding neural activity from the speech-specific brain areas, as pioneered by Guenther and colleagues (Guenther et al., 2009). This first requires to localize the speech-specific brain areas where to record neural activity, then to extract informative features, and finally decode it into control parameters for a speech synthesizer in order to generate speech in a closed-loop paradigm so that the subject can benefit the auditory feedback (**Fig. 1**).

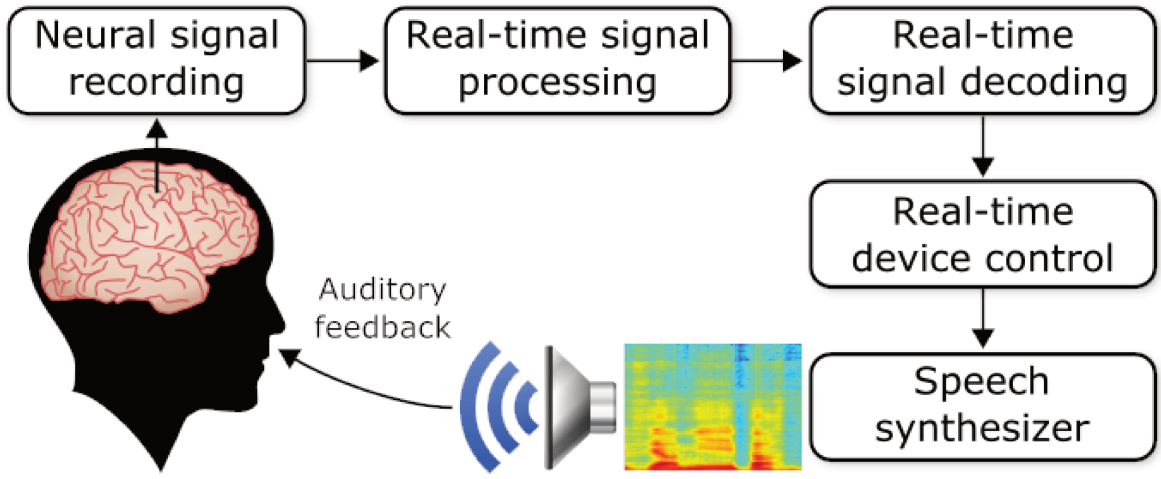


Fig. 1: Principle of a speech brain-computer interface. Neural activity is recorded from various speech-related brain areas and then processed to extract informative features, which are then decoded into control parameters for a speech synthesizer. The synthesis is performed in real-time so that the subject can benefit from the auditory feedback to better control the synthesizer.

In the following we will thus cover several findings that provide key information when considering speech restoration through a brain-computer interface. We will first briefly introduce the different types of neural signals that can be recorded and the different recording technologies available for BCIs. Then we will attempt to present the different brain areas that could be considered to record speech-specific neural activity. Finally, we will discuss the different decoding strategies that could be envisioned to predict speech from neural activity.

1. Neural activity recording

Even if neurons are not made of good conducting materials, elaborate mechanisms allow them to generate, transmit and maintain information under the form of electrical signals. Neurons are schematically made of three parts: the dendrites, the cell body or soma, and the axon (**Fig. 2**). The dendrites generally form a tree and receive inputs from the other nerve cells,

which are then integrated by the soma. The axon is the main conducting unit of the neuron and propagates signals to the other nerve cells.

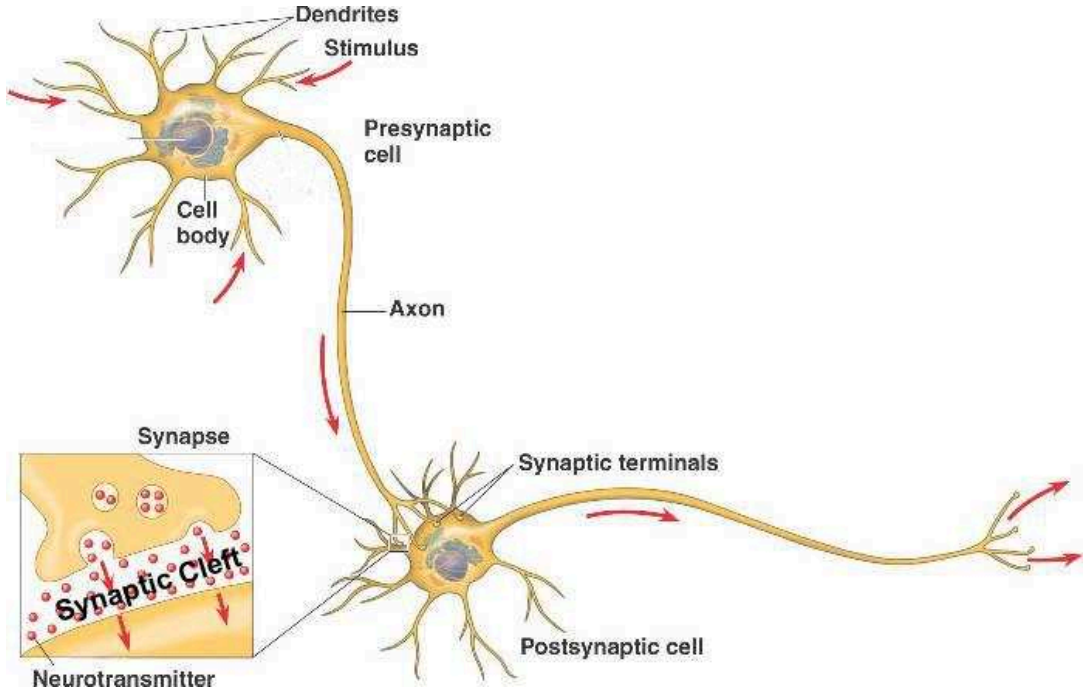


Fig. 2: Structure of a neuron. A neuron is composed by dendrites, the cell body and an axon that ends by synapses that contact other neurons. The transmission of action potentials at the synapses is generally chemical, by releasing neurotransmitters. Source: thatsbasicscience.blogspot.fr

The cellular membrane of neurons is sprinkled with ion pumps and leaky ion channels. While the membrane is an insulator and a diffusion barrier to the movements of ions – which are electrically charged particles, ion pumps actively push ions across the membrane and establish concentration gradients across the membrane. On the other hand, leaky ion channels passively allow or prevent specific ions from traveling through the cellular membrane down the concentration gradients. The difference in ions concentration gives rise to a difference in electric potential between the interior and the exterior of the cell, the transmembrane potential. At rest, there are concentration gradients of sodium and potassium ions across the cell membrane, with a higher concentration of sodium ions outside the neuron and a higher concentration of potassium ions inside the neuron. These gradients are maintained by sodium/potassium ion pumps which constantly push potassium in and sodium out the cell. The corresponding transmembrane potential is called resting potential. The resting potential is generally close to the potassium reversal potential, arround -70mV, meaning that the intra-cellular medium is more negatively charged than the extra-cellular medium.

A neuron typically receives input signals at the dendrites which are then spread through the soma. The axon of the other nerve cells contact the dendrites at sites called synapses (**Fig. 2**). The transmission of the neural signal at a synapse is generally chemical, through the release of neurotransmitters. In the case of an excitatory signal, these neurotransmitters open ligand-gated sodium channels, thus allowing sodium to flow into the cell, which increases the transmembrane potential. This flow of sodium ions travels toward the axon hillock, which is the part of the cell body that connects to the axon. Chemically generated synaptic currents are relatively slow phenomena of about 10 to 100 milliseconds. If the sum of all input currents is

high enough, an action potential (also called spike) is generated at the axon hillock and travels down the axon to the other nerve cells.

An action potential is essentially a short – about 3 milliseconds – auto-regenerating reversal of the transmembrane potential that propagates from the soma to the axon end (**Fig. 3**). Specific ion channels, called voltage-gated ion channels, are sensitive to the transmembrane potential and only open for a range of potential values. These channels are mostly concentrated at the axon hillock and open when the transmembrane potential increases to a certain threshold, typically about -55mV. When this threshold is reached, sodium voltage-gated channels open quickly, while potassium voltage-gated channels open more slowly, thus firstly giving rise to a sodium influx. This further increases the transmembrane potential, causing more channels to open. This exploding process goes on until all sodium channels are opened, reversing the polarity of the cell membrane, and is called depolarization. As the potential reaches its peak, sodium channels close while all potassium channels are opened, causing potassium ions to rush out of the cell and the potential to quickly decrease to its original resting value. This phase during which the potential decreases is called repolarization. Since potassium channels are also slow to close, potassium ions still leave the cells after reaching the resting potential, resulting in a negative overshoot before reaching the resting potential again, called hyperpolarization. During and shortly after an action potential, the part of the membrane that generated it is very difficult to stimulate to fire again. This period is called the refractory period.

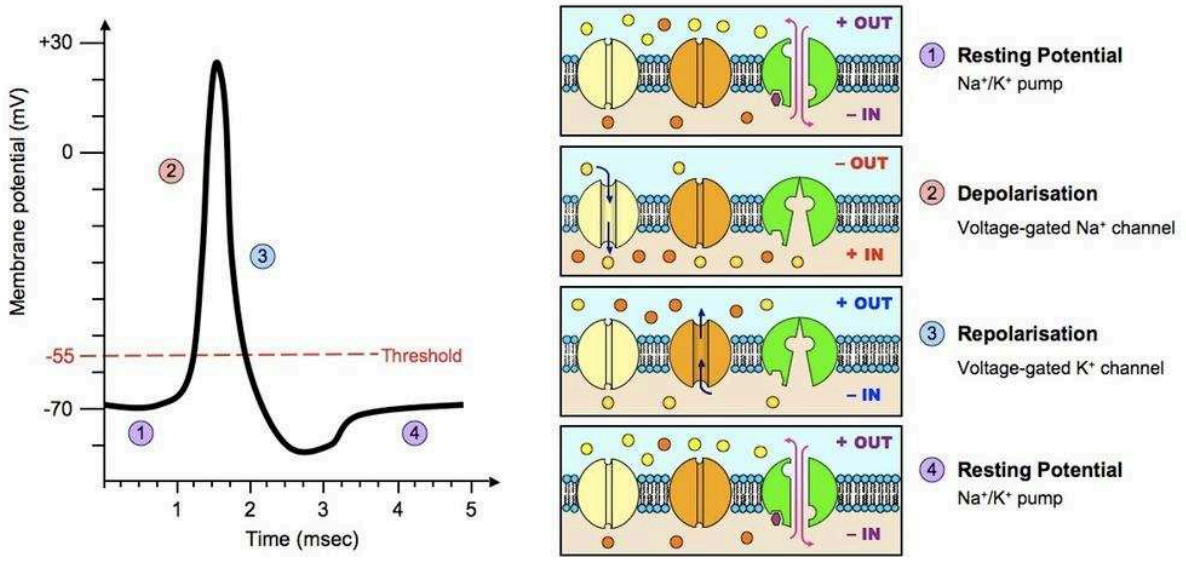


Fig. 3: Action potential. Left – An action potential is a short peak signal, that can be described by 4 phases: rest (1), depolarisation (2), repolarisation (3) and refractory period until rest (4). Right – The 4 phases of the action potential are generated by succesive activations and deactivations of ionic channels (green: Na^+/K^+ pump, light yellow: voltage-gated Na^+ channel, orange: Voltage-gated K^+ channel). Source: www.vce.bioninja.com.au.

The same mechanism is used to transmit and auto-regenerate the triggered action potentials along the axon. During an action potential, the influx of sodium ions at the basis of the axon spreads along the axon, which depolarizes the adjacent portion of axon, which in turn generates a similar action potential. Since the generation of an action potential induces a refractory period, the generated action potential propagates in only one direction. This self-induced process is repeated until the action potential reaches the end of the axon, at the synapses. At synapses, the arrival of an action potential can trigger the release of neuron transmitters, which will then in turn excite or inhibate other neurons, etc. While synaptic currents can be assimilated to dipoles

with an electrical amplitude decay inversely proportional to the squared distance to the source, action potentials can be assimilated to quadrupoles and thus decay much faster, with an amplitude inversely proportional to the square of the squared distance to the source. Thus, measuring an individual action potential requires to be relatively close to its source, while measuring synaptic currents can be done on a much greater distance.

All these processes are energy consuming, and active neural cells have a higher demand of energy in the form of oxygen or glucose. Thus neural activity consists not only in transmembrane ions currents and changes in electric potentials but as well in variations of supply in oxygen or glucose. Neural activity can thus be recorded by measuring the latter, i.e. the metabolic signals, or by measuring the electric currents and potentials, i.e. the electrophysiological signals.

2. Metabolic signals recording

Active neurons have a higher demand of energy in the form of oxygen and glucose, which results in an increased blood flow in the active brain areas (Logothetis et al., 2001). Thus one way of measuring the brain activity is to measure the flow of oxygenated blood that travels through the vessels of the brain, or the variations of concentration in oxygen or glucose. Blood vessels form a dense network in the brain so that relatively low scale changes in blood flow can be measured to reflect the neural activity of a small brain area, thus offering possibilities for good spatial resolution. However, this metabolic process is quite slow (about a second) which clearly limits the temporal resolution of the methods that exploit this phenomenum. Several methods take advantage of changes in blood oxygenation or glucose concentration to measure the brain activity: functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), optical imaging of intrinsic signals (OIIS) and positron emission tomography (PET).

a. Functional magnetic resonance imaging

fMRI uses a strong magnetic field to detect changes of the magnetic properties associated with blood flow. The most commonly used form of fMRI is the blood-oxygen-level dependent fMRI (BOLD-fMRI) which detects changes of blood oxygenation. Indeed, when a specific region of the cortex increases its activity in response to a task, the extraction of oxygen from the local capillaries leads to an initial drop in oxygenated hemoglobin (also called oxyhemoglobin), which is then followed – after about a second – by a deliver of a surplus of oxygenated hemoglobin that washes away deoxygenated hemoglobin (also called deoxyhemoglobin). There is thus a change in the relative levels of oxyhemoglobin and deoxyhemoglobin that can be detected given that they have different magnetic susceptibility, which allows to quantify the blood oxygenation level (Ogawa et al., 1990). fMRI is non-invasive and offers a relatively high spatial resolution (about a millimeter). However, the two main drawbacks of this technique, in addition to its high cost, are its low temporal resolution and the fact that it is not portable.

b. Functional near-infrared spectroscopy

fNIRS also takes advantages of differences between oxygenated and deoxygenated hemoglobin by using near-infrared imaging. Indeed, while the skin, tissue and bones have a good transparency to near-infrared light – wavelength from 700 to 900 nanometers (Jobsis, 1977), oxygenated hemoglobin and deoxygenated hemoglobin strongly absorb it, each having a specific absorption spectra. By combining near-infrared light at different wavelengths and measuring the light attenuation, the relative concentrations of oxyhemoglobin and deoxyhemoglobin can be estimated, thus allowing an indirect measure of the brain activity. As opposed to fMRI, fNIRS has a lower cost, and is portable even in freely-moving subjects. However, it suffers as well from a poor temporal resolution and has a lower signal-to-noise ratio and spatial resolution than fMRI (Cui et al., 2011).

c. Optical imaging of intrinsic signals

OIIS maps the functional cortical activity by detecting changes in cortical light reflectance that are related to changes in neural activity (Grinvald and Bonhoeffer, 1999). Indeed, similarly to fNIRS, changes in blood volume and ratio of oxygenated hemoglobin and deoxygenated hemoglobin induce changes in the cortical light reflectance. Moreover, the light reflectance is also affected by light scattering which is tightly coupled, spatially and temporally, with neural activity. Indeed, when optically imaging the living brain, the incident light is scattered to some extent as it penetrates and is reflected through the neural tissue. This light scattering increases when the neural activity increases, which is thought to result from ion and water movement, expansion and contraction of extracellular spaces, capillary expansion or neurotransmitter release (Cohen, 1973). These changes can be detected by using a charged-coupled device camera that captures images of the exposed cortex both at rest and during activity. Different wavelengths are used to measure the different signal components that affect the light reflectance, and near-infrared wavelengths are generally used (Zepeda et al., 2004). OIIS offers high spatial resolution (about 100 microns), is relatively non-invasive, and can be used chronically (several weeks or months) in the same subject. However, its temporal resolution is limited by the nature of the metabolical signals being recorded (about half a second).

d. Positron emission tomography

As opposed to the other methods, PET requires the injection of chemicals in the bloodstream to measure brain activity. The chemicals are radioactive tracers that are radioactively labeled so that their radioactive emissions can be measured, and are chosen among chemicals involved in the metabolic processes of brain activity, especially glucose, allowing to indirectly measure the flow of blood to different parts of the brain. Other PET chemicals have also been developped that are ligands for specific types or neuroreceptors, making this technique particularly advantageous for studying specific diseases. As fMRI and fNIRS, PET is a non-invasive technique and has been already used for in-vivo recordings. Moreover, it offers relatively good temporal (about 0.2 seconds) and spatial (about 1mm) resolutions (Castermans et al., 2014). However, PET is non-portable and requires the injection of radioactive tracers that have a quick decay of radioactivity, limiting its usage to short tasks.

3. Electrophysiological signals recording

As opposed to metabolic methods, electrophysiological approaches directly record the neural electric signals in the intracellular or extracellular medium. Here we will only consider extracellular recordings that are more suitable for BCIs since intracellular recording require high-precision experimental setup generally incompatible with a freely-moving subject. The extracellular signal reflects the sum of synaptic currents, action potentials, flow of the ions through the membrane channels, etc. It is thus a combination of numerous phenomena, some being slow, other being fasts, and with varying amplitudes – principally because of the distance that exists between the recording and the source sites. This brain signal can be measured using electrodes that measure variations of the electric potential or magnetic field in their surrounding (**Fig. 4**). Most commonly used methods are magnetoencephalography (MEG), electroencephalography (EEG), electrocorticography (ECoG), micro-electrocorticography (µECoG), stereo-electroencephalography (SEEG) and using intracortical micro-electrodes arrays (MEAs). MEG measures changes in the magnetic fields induced by electrical currents using magnetic sensors placed arround the head. EEG consists in measuring changes in the electrical potential by large electrodes placed over the scalp. ECoG is when medium-sized electrodes are placed directly over the cortex – above or under the duramea, and µECoG corresponds to high-density ECoG. SEEG consists in implanting in depth thin wires on which are spaced several electrodes through some openings in the skull. Finally, MEAs are small grids of micro-sized penetrating electrodes placed in the cortex.

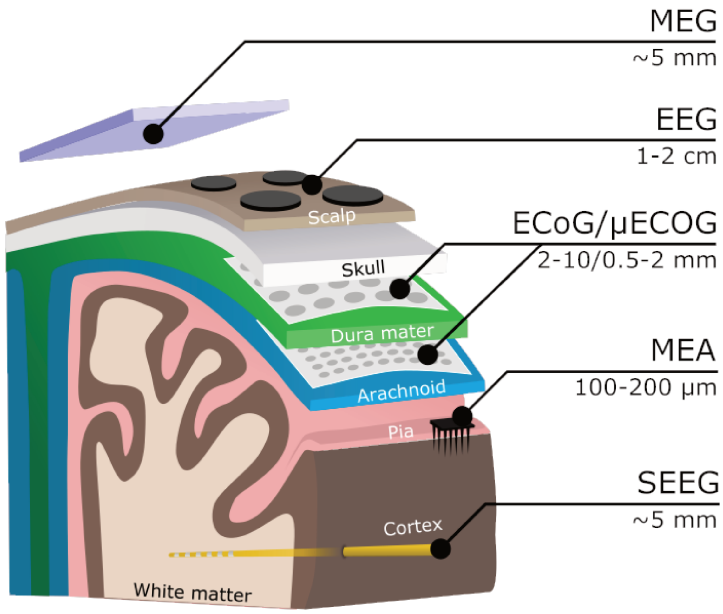


Fig. 4: MEG, EEG, ECoG, µECoG, SEEG and MEA recordings. With MEG, magnetic sensors are placed all around the head. In EEG, relatively large electrodes are placed over the scalp. In ECoG, smaller electrodes are placed under the skull, either above (epidural) or under (subdural) the dura mater. µECoG is similar to ECoG but uses very high density grids with smaller electrodes. SEEG are thin wires on which are spaced several electrodes that penetrates the brain in depth. MEA consists of micro-electrodes penetrating the cortex. Adapted from (Jorfi et al., 2015).

a. Electroencephalography

Since neural activity is mostly made of ion currents and variations of potentials, measuring the brain activity can be achieved by measuring changes in difference of potential between

several electrodes – i.e. some electrical conductor – that are placed over the regions of interest, relatively to some reference potential, using a voltmeter. For this purpose, EEG is one of the most widely used methods for measuring the brain electric activity, particularly because it has a very low cost, is portable and non-invasive. Indeed, in EEG, relatively large electrodes, generally distributed on a helmet, are placed over the scalp, without the need for surgery. However, because of the distance between the recording electrodes and the current sources in the brain, and because of the differences in conductivity between the brain and the skull, the measured signal is generally distorted and attenuated, thus making it difficult to directly relate the EEG with the electrical activity of individual neurons. Thus, EEG mostly reflects the summation of the synchronous activity of thousands to millions of neurons that have similar orientation. The measured signal mostly reflects synaptic currents since EEG electrodes are too far from the neurons to measure action potentials that decay much faster with the distance to their source. Synaptic currents can be assimilated to dipoles, and are summed when these dipoles have similar orientation. The EEG signals is therefore mostly thought to come from the activity of pyramidal neurons since these neurons are well-aligned and tend to fire together. The spatial resolution of EEG is thus relatively low (about a centimeter) when compared to other recording techniques, even if there are EEG arrays offering higher spatial density of electrodes (Tucker, 1993).

b. Magnetoencephalography

According to the electromagnetic laws, varying electrical currents induce varying magnetic fields. Thus, a way to record the brain activity is to record the variations of the magnetic field that surrounds the brain using magnetic sensors. This is generally achieved using magnetoencephalography (MEG). MEG generally uses magnetometers made of several supraconductor loops, called superconducting quantum interference devices or SQUIDs, to measure the magnetic field surrounding the subject's head induced by the neuronal currents (Hämäläinen et al., 1993). This non-invasive method has a good temporal resolution (of the order of the millisecond) and a good spatial resolution (of the order of several millimeters). One main advantage of measuring the magnetic field with regards to EEG is that it does not suffer from the distortions mostly caused by the difference of conductivity of the skull. However, MEG remains a very costly method – mostly because of the cooling of the supraconductors – and is not portable since it has to be used in a magnetic shielded room, which makes it not suitable for BCI applications. Nonetheless, other types of magnetic sensors are being envisioned to monitor brain activity, such as optical pumping magnetometers (Lembke et al., 2014), that could overcome some of the current MEG issues, especially the need for cooling.

c. Electrocorticography

ECoG recordings are becoming more and more widespread to study cortical phenomena in clinical conditions (Engel et al., 2005). Neural electrical activity is directly recorded at the surface of the cortex using arrays of metallic electrodes – generally made of platinium or iridium. Althought this is an invasive method that requires to open the skull, it allows to avoid distortion effects mostly caused by the difference of conductivity of the skull. This method offers both a good temporal resolution (of the order of milliseconds) and a good spatial

resolution (of the order of several millimeters, according to the electrodes size and spacing). Tipical ECoG grids contains several dozens of electrodes which allow to cover large cortical areas (of the order of 100cm²). ECoG signals can be decomposed into several frequency bands of special interest (Buzsáki and Draguhn, 2004) named according to their order of discovery: delta (1-4Hz), theta (4-10Hz), alpha (7-12Hz), beta (10-30Hz), gamma (30-80Hz) and high-gamma (80-200Hz). For instance, fast rythms, such as high-gamma activity, have been shown to be linked to cognitive processing and correlated to the firing rate of action potentials of the neurons (Ray and Maunsell, 2011).

d. Micro-electrocorticography

Some ECoG grids can contain up to several hundreds of electrodes, that are smaller than normal ECoG electrodes. Using such grids is generally referred to as micro-electrocorticography (µECoG) and allows to significantly increase the spatial resolution with regards to using normal ECoG grids, reaching the submillimeter scale. This allows to record the local field potential (LFP) at the surface of the cortex. The LFP is composed by all potential fluctuations in a small volume, and is generally low-pass filtered below 300Hz, and thus does not usually contain individual action potentials. Synaptic currents largely contribute to the LFP since they are slow events that can easily overlap.

e. Stereo-electroencephalography

The LFP can also be recorded by inserting electrodes inside the brain. Stereo-electroencephalography (SEEG) uses thin wires along which are spaced several macroscopic electrodes that are inserted in the brain through small holes pierced in the skull. This intracortical method offers a spatial resolution comparable to ECoG (about several millimeters) but allows to probe deep regions of the brain (up to several centimeters under the cortical surface). SEEG is slightly less invasive than ECoG and reduces risks of infection since it does not require to open the skull but instead to pierce a hole at the insertion location. However, SEEG up to 20 wires are generally necessary to cover sufficient brain volume since each wire has few electrodes (typically about 5), and the exact positioning of each electrode is difficult to achieve since the implantation trajectory must avoid blood vessels.

f. Intracortical micro-electrodes arrays

Inserting microscopic electrodes – called micro-electrodes – in the brain allows not only to record the LFP but individual action potentials as well. These electrodes can be individually inserted, sometimes to depth of several millimeters, or organised into small grids – called micro-electrode arrays (MEAs) – that penetrate the brain over one or two millimeters in order to reach specific neuron layers of the cortex. Using high-density intracortical MEAs allows to record the action potentials of the neurons in a small volume (Kipke et al., 2008). This method offers both high temporal resolution and high spatial resolution (of the order of 100 micrometers). However, MEAs only allow to cover small cortical areas (of the order of 1cm²), which can be slightly compensated by simultaneously using several electrode arrays. The high temporal and spatial resolutions of this method allow to record individual action potentials, which form varies

according to the type of neuron that triggered it, and to the position of the electrodes with regards to the cell, and other factors (Gold et al., 2006). This allows to automatically detect and classify action potential according to the neuron that triggered it. The high-frequency signal containing action potentials from several neurons is generally referred to as multiple-unit activity, as opposition to the isolated neuron spikes which are referred to as single-unit activity.

II. Cortical speech production areas

With brain-computer interfaces, neural activity is recorded in order to be decoded into control parameters for an effector. Recording brain activity that is somewhat correlated to the task that the subject wants to perform seems to be the best strategy in order to maximize the decoding accuracy and reduce the cognitive load for the user. For a speech BCI, speech-related brain activity must thus be recorded and a choice needs to be made on the cortical areas to record and decode activity from.

1. Speech areas

Speech processing by the human brain involves a wide cortical network, which has been modeled by two main information streams linking auditory areas of the superior temporal plane to articulatory areas of frontal regions, one ventral and the other dorsal (Hickok and Poeppel, 2004, 2007). The ventral stream involves regions of the middle and inferior temporal lobe and maps speech sounds to meaning, while the dorsal stream runs through the dorsal part of the posterior temporal lobe at the temporo-parietal junction and is responsible for the sensori-motor integration of speech by mapping speech sounds to articulatory representations (Friederici, 2011; Hickok et al., 2011). Lesions of ventral stream regions of the temporal lobe result in Wernicke aphasia characterized by impairments of speech comprehension, while lesions of frontal areas result in Broca aphasia characterized by impairments of speech production. Classically, the dorsal stream has been described to be largely left-hemisphere dominant, but several studies indicate that many aspects of speech production activate cortical areas of the dorsal stream bilaterally (Pulvermüller et al., 2006; Cogan et al., 2014; Geranmayeh et al., 2014; Keller and Kell, 2016). Thus, several cortical areas can be considered in order to decode speech from neural activity.

One possibility could be to record neural activity from the auditory areas, which are known to encode parts of the spectro-temporal representation of perceived sounds, including speech (Engineer et al., 2008; Mesgarani et al., 2008; Steinschneider et al., 2013). Indeed, some works suggest that these areas could as well be involved in perceiving inner imagined speech, or covert speech (Pei et al., 2011a; Martin et al., 2014). However, these areas are not specific to self-produced speech, but rather to all the sounds a person is exposed to, including self-produced and inner speech, as well as other people's speech, and non-speech environmental sounds. Using brain activity from these areas for a speech BCI could lead to difficulties when decoding only self speech intention. For this reason, it seems more relevant to probe neural activity from brain areas specifically involved in speech production.

2. Speech production

Speech production can take various forms, and can be mainly divided between overt speech and covert speech. While overt speech is when sounds are pronounced out loud using one's vocal apparatus, covert speech is an inner phenomenon, from one's mind to itself, without any production of sound or other exterior phenomena. Covert speech is of particular interest for a speech BCI, since a speech BCI is meant to be used by subjects that are not able to produce overt speech, such as locked-in patients. Some studies already investigated neural differences between overt and covert speech. However, as pointed by (Perrone-Bertolotti et al., 2014), one can consider many different types of covert speech – from inner thoughts to voluntary imagination of the acoustic of a speech sound – and as many different types of overt and covert speech tasks as there are studies – from word repetition to object denomination or text reading, making it difficult to compare one study with another.

a. Overt speech

The production of overt speech implies movements of the speech articulators (tongue, lips, jaw, velum and to a certain extent the lungs and vocal folds involved in phonation). It is thus expected that motor areas of the brain are largely activated during overt speech. Activity in the left primary motor and premotor cortices during overt speech has been indeed largely reported (Kellis et al., 2010; Pei et al., 2011a; Herff et al., 2015; Lotte et al., 2015), and precedes speech production by several dozens of milliseconds (Pei et al., 2011b; Herff et al., 2015). In particular, speech production is classically associated with a decrease of signal power in the beta frequency range and usually an increase in the high gamma frequency range over temporal and motor frontal areas (Canolty et al., 2007; Pei et al., 2011b; Toyoda et al., 2014) althought gamma attenuation was observed in more anterior frontal speech cortex including Broca area (Lachaux et al., 2008; Wu et al., 2011; Toyoda et al., 2014).

Using µECoG to record local field potentials (LFPs) on the surface of the face motor cortex (FMC, Brodmann's area 4 and 6) and Wernicke's area – which is considered as an area for language comprehension, it has been observed that the FMC was predominantly activated during a word repetition task and exhibited frequency features aligned with individual words, but that it was less activated during a conversational task, while the opposite was true for Wernicke's area (Kellis et al., 2010). Gamma-band activity from the FMC was further shown to be the most informative signal when decoding spoken words (Mugler et al., 2014), which confirms the major role of the FMC in overt speech production.

Moreover, it has been shown that the ventral half of the lateral sensorimotor cortex (vSMC) shows significant activity increase during movements of the speech articulators and thus during overt speech (Brown et al., 2008, 2009; Grabski et al., 2012; Bouchard et al., 2013). The vSMC is composed of the pre- and post-central gyri (Brodmann areas 1, 2, 3, 4 and 6), and the gyral area directly ventral to the termination of the central sulcus (Brodmann area 43). The activation of the vSMC during speech production has been detailed using µECoG recordings in (Bouchard et al., 2013). In this study, a classification-based analysis was conducted and exhibited a somatotopic organization of the vSMC by speech articulator (**Fig. 5**). Four speech articulators were considered (lips, jaw, tongue and larynx). Some single electrodes showed a clear tuning preference to individual articulators and some single electrodes had functional representation

of multiple articulators. Moreover, cortical representations exhibited a hierarchical organization according to the articulatory properties of the phonemes.

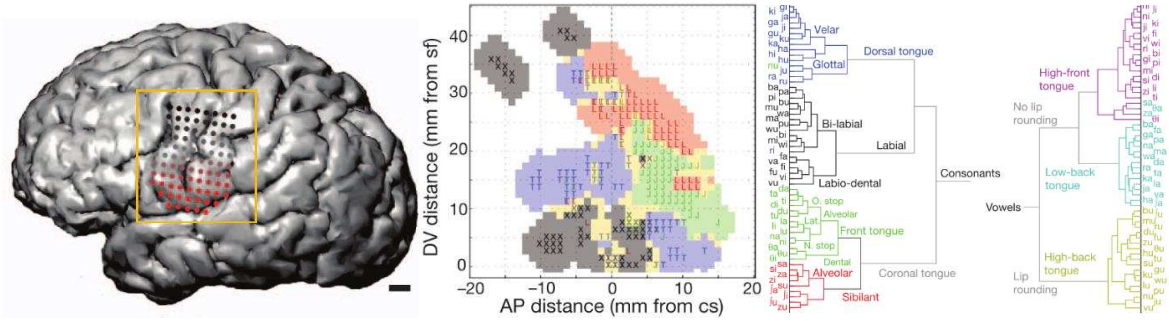


Fig. 5: Spatial organization of the vSMC during speech production. Left – Location of the ECoG grid electrodes over the vSMC. Middle – Functional somatotopic organization of speech-articulator representations in vSMC plotted with regard to the anteroposterior (AP) distance from the central sulcus and dorsoventral (DV) distance from the Sylvian fissure. Lips (L, red); jaw (J, green); tongue (T, blue); larynx (X, black); mixed (yellow). Right – Hierarchical clustering of the cortical activities for consonants (left) and vowels (right), with branches labeled with linguistic categories. (Bouchard et al., 2013).

It was further shown recently that during speech production, the activity of the speech sensorimotor cortex – including the vSMC – is tuned (i.e. it is modulated and specific) to the articulatory properties of the produced sounds but not to their acoustic properties (Cheung et al., 2016).

While speech production was originally thought as being predominantly left-lateralized, several studies reported bilateral activity (Petersen et al., 1988; Palmer et al., 2001; Cogan et al., 2014; Martin et al., 2014). Continuous production of narrative speech was also shown to activate frontal motor speech regions as well as comprehension temporal and parietal areas bilaterally (Silbert et al., 2014). Intraoperative functional mapping data collected in a high number of patients undergoing awake surgery also reported bilateral critical motor and premotor regions for overt speech production (Tate et al., 2014). The right hemisphere is also clearly activated during synchronized speaking (e.g. singing with a group of people) in several regions including the temporal pole, inferior frontal gyrus, and supramarginal gyrus (Jasmin et al., 2016).

When more complex tasks are considered that require additional semantic, lexical, or phonological processing, specific activations are observed in the left inferior frontal cortex (Petersen et al., 1988, 1989; Price et al., 1994; Sörös et al., 2006; Basho et al., 2007). These findings suggest that speech production becomes left lateralized when inner high-level processing is required.

b. Covert speech

Covert speech brain activity was originally envisioned as overt speech activity without motor activity. However, there are physiological evidences of motor activity during covert speech: for instance, an increase in the electromyographic (EMG) activity has been reported in some of the lips muscles during covert auditory hallucinations in patients with schizophrenia, that was not caused by a general increase in muscular tension (Rapin et al., 2013). Another study, showed that repetitive transcranial magnetic stimulation (rTMS) of left motor hemisphere frontal sites perturbed covert speech, resulting in longer latencies to perform a syllable counting task (Aziz-Zadeh et al., 2005).

In general, covert speech has been found to activate similar brain areas but with a lesser amplitude than overt speech across most ventral and dorsal stream areas (Price et al., 1994; Ryding et al., 1996; Palmer et al., 2001; Shuster and Lemieux, 2005). In particular, as for high-level overt speech production, cortical activity underlying covert speech production is left lateralized with strong activation of the left motor, premotor and inferior frontal cortex (Ryding et al., 1996; Palmer et al., 2001; Keller and Kell, 2016).

An fMRI study of overt and covert naming of visually presented letters or animal names starting by the presented letter showed that significant activation in Broca's area (Brodmann's areas 44 and 45) was detected during both overt and covert speech (Huang et al., 2001). This suggests that if Broca's area plays a role in phonological or articulatory coding, this role is not particular to overt production, that is, it is not tied specifically to motor output. They however showed that the face motor cortex was only activated during overt speech production, which seems in contradiction with other studies.

Indeed, several studies suggested that the primary motor cortex activity contains informative content to decode covert speech (Pei et al., 2011a; Martin et al., 2014). In (Martin et al., 2014), several subjects had to read a text both overtly and covertly while ECoG grids were used to record the corresponding brain activity at multiple sites. A model was then built on the overt speech data in order to predict acoustic features from the brain activity, and the same model aplied to covert speech data led to results over chance level, suggesting that overt and covert speech share a part of their neural substrate. In particular, these results showed that there was no significant change of activity over the vSMC between overt and covert speech production. In (Pei et al., 2011a), informative areas for decoding covert consonants and vowels were distinguished. In this study, results from all the subjects were combined to identify the most informative areas for decoding phones. For consonants, the most informative areas were located in a temporal region near Wernicke's area, while the most informative areas for vowels were located in the primary motor cortex. This suggests that covert word repetition consists both in imagining the perceptual qualities and the processes that simulate the motor actions necessary for speech production.

Overall, these results suggest that the left inferior frontal region encompassing Brodmann areas 4, 6, 43, 44 and 45, are relevant candidates from which to probe and decode neural activity for the control of a speech BCI. However, these studies also show that there is a large variability in the exact localization of the relevant areas, probably due to individual specificities (morphology, somatotopy, …) or differences in the task being performed, or event in the method used to record the neural activity. Thus, while there are evidences of cortical activation in these areas during covert speech, these results also suggest that speech production areas must be identified individually, for each subject.

III. Speech decoding from neural activity

The different studies that aimed at decoding speech features from neural activity can be mainly divided into two categories: those using a discrete decoding approach, and those using a continuous decoding approach.

1. Discrete decoding

Discrete decoding approaches aim at classifying the neural activity into several categories, generally corresponding to phonetic units such as phones (Brumberg et al., 2011; Pei et al., 2011a; Tankus et al., 2012; Mugler et al., 2014; Song et al., 2014) or full words (Kellis et al., 2010). Such information can then be used to synthesize speech, for instance using classical text-to-speech synthesis. In (Mugler et al., 2014), on average 20% of accuracy was achieved at decoding 31 different English phones from ECoG data recorded from 4 subjects overtly reading isolated monosyllabic words, with up to 36% of accuracy in one subject, using 6 electrodes located over the ventral somatosensory region. In (Pei et al., 2011a) phones were decoded with similar accuracy, not only from overt speech but also from covert speech. Indeed, to a lesser extent, ECoG data could also be used to predict covertly imagined speech not actually overtly pronounced by the subject (Pei et al., 2011a). The prediction of covert speech was in general more limited than for overt speech but above chance level, and more reliable for vowels than for consonants. Using micro-electrodes in a patient suffering from locked-in syndrom and thus who could not produce any overt movement nor speech, 38 different American English phones could be decoded with about 20% accuracy (Brumberg et al., 2011). In this study, and as opposed to (Pei et al., 2011a), consonants were found to be more reliably decoded than vowels. By analyzing the articulatory properties of each classified phone, the authors found out that the implant might be located in an area representing lip movements. The positioning of the micro-electrodes was optimized prior to surgery using fMRI to determine the locations of the brain areas active during speech production attempts, which allowed to localize a single area on the left precentral gyrus, lying on or near the border between pre-motor and primary motor cortex. In (Kellis et al., 2010), about 50% accuracy was achieved at decoding full words among a subset of ten different words from neural activity recorded during overt speech using micro-electrodes implanted over the face motor cortex. While this is an encouraging result, direct classification of full words does not seem adequate for a speech BCI given the large dictionary size needed to represent all words and phrases used in conversational speech – generally about 3,000 words. On the other hand, phonemes are the smallest units from which speech is built, and additional linguistic knowledge can be taken into account to improve the decoding of full sentences, as it is classically done in automatic speech recognition. In (Herff et al., 2015), such linguistic knowledge was added by limiting the vocabulary to a predefined dictionary, and by using a statistical language model that helped predicting a word given the preceding one. While the raw phone recognition accuracy ranged from about 10% to 50% according to the subject and session for a total of 20 different English phones in an overt reading task, limiting the vocabulary to 10 words and using a statistical language model computed from the read texts allowed to reach a word recognition accuracy of about 75%.

2. Continuous decoding

By contrast with discrete classifiers, continuous decoding approaches do not rely on decoding an intermediate discrete representation – such as a phonemes. Instead, they directly infer continuous parameters – generally acoustic trajectories – from the neural activity. Such acoustic trajectories can then be used by a parametric speech synthesizer to synthesize speech (Guenther et al., 2009). While in discrete decoding approaches neural data can typically be decoded only after a full speech segment – phone, word or sentence – has been pronounced or imagined, continuous approaches rather directly predict speech typically in a frame-by-frame fashion. Very few studies considered continuous decoding of produced or intended speech. In (Martin et al., 2014) a spectro-temporal representation of sounds (obtained using wavelet transform on the audio signal with 32 logarithmically-spaced frequency bins between 180Hz and 7kHz) was directly inferred from ECoG recordings, both during overt and covert speech. Although this approach could not produce intelligible speech, the overall time-frequency structure of the speech spectrograms could be well estimated. Moreover, most informative electrodes for decoding covert speech were mostly located in the vSMC. In (Guenther et al., 2009), the activity recorded from intracortical electrodes localized in motor part of the vSMC was used by a locked-in patient to control in real-time a speech synthesizer in order to produce vowels and transitions between them. The electrode recorded action potentials from individual neurons, which could be automatically classified according to the emitter cell, in order to compute the firing rate of each recorded neuron over time. This firing rate was computed in real-time in order to infer the frequencies of the two first formants of the speech signal, which are the two first major peaks of the speech spectrum envelope. Although vowels are plainly characterized by the position and amplitude of their 4 or 5 formants, using only the position of these two first formants is enough to distinguish most of them.

IV. Conclusion

1. Choice of a recording technique to monitor speech brain signals

In this chapter, we presented different techniques to record neural activity: functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS) , optical imaging of intrinsic signals (OIIS), positron emission tomography (PET), magnetoencephalography (MEG), electroencephalography (EEG), electrocorticography (ECoG), micro-electrocorticography (µECoG), stereo-electroencephalography (SEEG) and intracortical micro-electrodes arrays (MEAs). The following table (**Table 1**) resumes the properties of each of these methods:

Recording method	Neural signal type	Temporal Resolution	Spatial Resolution	Invasiveness	Portability
fMRI	Metabolic	~1 s	~1 mm	Non-invasive	Non-portable
fNIRS	Metabolic	~1 s	~2 cm	Non-invasive	Portable
OIIS	Metabolic	~0.5 s	~0.1 mm	Slightly-Invasive	Portable
PET	Metabolic	~0.2 s	~1 mm	Non-invasive	Non-portable
MEG	Magnetic	~0.05 s	~5 mm	Non-invasive	Non-portable
EEG	Electrical	~0.001 s	~1 cm	Non-invasive	Portable
ECoG	Electrical	~0.001 s	~5 mm	Invasive	Portable
µECOG	Electrical	~0.001 s	~0.5 mm	Invasive	Portable
SEEG	Electrical	~0.001 s	~5 mm	Invasive	Portable
MEAs	Electrical	~0.001 s	~0.1 mm	Invasive	Portable

Table 1: Comparison of different neural activity recording methods. Colors are indicative and reflect the compliance of the method with regard to each criteria for BCI in the context of speech rehabilitation.

Althought MEG and fMRI have already been considered for decoding neural activity (Weiskopf et al., 2004; Mellinger et al., 2007; Formisano et al., 2008; Waldert et al., 2008; Lee et al., 2011; Quandt et al., 2012; Choupan et al., 2013) including speech-related activity (Formisano et al., 2008; Koskinen et al., 2013; Bonte et al., 2014; Correia et al., 2014, 2015), they do not seem suitable for BCI applications. Indeed, they both require expensive equipments, do not work in freely moving subject, and are not portable. Moreover, fMRI, as well as PET and OIIS, have a poor temporal resolution directly linked to the nature of the metabolic signals being recorded.

While fNIRS is portable and has already been used for decoding speech neural activity and BCI applications (Coyle et al., 2007; Sitaram et al., 2007; Herff et al., 2012), it as well suffers from a poor temporal resolution of several seconds. However, speech is made of fast events, a single phone lasting from about 10 to 100 milliseconds. Thus, fNIRS does not seem suitable for a speech BCI in which speech is continuously decoded from the neural activity.

Even if EEG has been widely used for decoding brain activity and BCI applications for communication purposes (Hinterberger et al., 2003; Deng et al., 2010; Brandmeyer et al., 2013; Grau et al., 2014; Yoshimura et al., 2016), it is mostly limited to the classification of the neural activity into a few categories, generally 2 or 3. This limitation mostly originates in its poor spatial resolution that prevents to track ongoing neural activity with sufficient details to enable, for instance, the prediction of continuous intelligible speech from brain signals. While this might be sufficient to select a letter on a screen or move a computer mouse cursor, it does not seem suitable for a continuous decoding of speech. Moreover, EEG is very prune to motion artifacts, especially from head movements, mucle activity and environmental electromagnetic noise.

SEEG has been successfully used to understand the cortical dynamics underlying speech and language perception (Liegeois-Chauvel et al., 1999; Basirat et al., 2008; Sahin et al., 2009; Fontolan et al., 2014). In particular it has helped to highlight how brain oscillations encode the rhythmic properties of speech, with a strong coupling of the theta rhythm to the tempo of syllables occurrence in speech, and associated nested modulation of gamma-band signals possibly encoding transient acoustic speech features (Giraud and Poeppel, 2012; Morillon et al., 2012). However, the limited spatial coverage of SEEG precludes the access to the detailed dynamics of frontal motor speech areas and may limit the possibility to decode with sufficient details a continuous speech flow produced either overtly or covertly.

On the other hand, ECoG has been largely used with success to decode above chance level speech intention (Kanas et al., 2014), place and manner of articulation (Lotte et al., 2015), phones (Pei et al., 2011a; Mugler et al., 2014; Song et al., 2014), words (Herff et al., 2015) or even full sentences (Herff et al., 2015). Similarly, µECoG recordings have been used to decode words froms LFPs (Kellis et al., 2010). Micro-electrode recordings have been used to decode phones from spiking activity (Brumberg et al., 2011), with accuracy superior or comparable to ECoG studies. While ECoG recordings have been more studied than spiking activity for decoding speech features, in other BCI fields, such as motor rehabilitation, spiking activity generally allows the decoding of more degrees of freedom – with latest studies showing the control of devices with up to ten degrees of freedom (Collinger et al., 2013; Wodlinger et al., 2014) which is about the same as the number of degrees of freedom our vocal apparatus uses when producing speech (Beautemps et al., 2001), with relatively good accuracy. Moreover, spiking activity recorded from the speech motor cortex was even used to control a speech synthesizer in real-time to continuously produce vowels (Guenther et al., 2009). Another study showed that some medial-frontal neurons had very specific tuning to individual vowels (Tankus et al., 2012). This suggests that high-density recording devices, such as micro-electrode arrays (MEAs), are needed in order to capture the fine features of neural activity underlying speech production and that higher performance is likely to be expected from denser recordings.

2. Choice of a brain region

In this chapter, we also briefly presented the different cortical areas involved during speech, and more especially during production of both overt and covert speech. In particular, there are several evidences of activation of frontal areas – especially Brodmann areas 4, 6, 43, 44 and 45 – during speech production, including imagined speech. However, there was a high variability

in the identified areas across studies and even across individuals in the same study. This suggests that speech production areas must be identified individually, for each subject.

Moreover, it should be noted that although aphasia caused by strokes very often implies the speech motor cortex or other cortical areas necessary for speech production, this is not the case for other types of aphasia, such as in locked-in patients or patients with ALS, for whom cortical speech activity can be intact or largely conserved and thus exploitable in a BCI perspective.

We previously motivated our choice for using MEAs to record neural activity. However, such MEAs are usually small and can only cover a small surface of the brain. On the other hand, ECoG grid can cover larger cortical areas. A possible way to take advantage of both approaches could be to perform first ECoG recordings in order to individually identify the speech areas, then use this information to optimize the positioning of one or several MEAs.

3. Choice of a decoding and synthesis approach

In this chapter we presented two main approaches for decoding speech from neural activity: discrete and continuous decoding approach. Both approaches showed very promising results. While discrete approaches can take advantage of linguistic knowledge to improve decoding performance (Herff et al., 2015), this comes at the price of an additional delay between speech intention, and actual speech synthesis since recognition of complete speech segments – from phones to sentences – is generally required. On the other hand, continuous approaches can perform the synthesis on a frame-by-frame basis and provide an almost instantaneous feedback to the subject (Guenther et al., 2009). This is a potential asset for continuously predicting speech at a natural pace since it has been shown that feedback delays above 50ms generally disturbs speech production (Lincoln et al., 2006). Moreover, several BCI studies pointed out the great importance of subject's training in improving their control accuracy (Ganguly and Carmena, 2009; Wodlinger et al., 2014). It is more likely that a direct feedback would improve training with respect to a delayed feedback, since it could allow the subjects to directly compensate during an ongoing speech production.

For these reasons, we made the choice to consider continuous speech decoding in the present work. However, the existing studies that continuously decoded speech from neural activity generally considered the decoding of acoustic features – such as the speech spectrum (Martin et al., 2014) or formant trajectories (Guenther et al., 2009) – while recent work showed that the frontal speech motor cortical regions are rather tuned to the articulatory than to the acoustic properties of speech (Bouchard et al., 2013; Cheung et al., 2016). While neural data could be decoded directly into acoustic parameters, these data support the hypothesis that a relevant strategy could be to consider a more "indirect" approach accounting for the articulatory activity of the vocal tract under control of the speech sensorimotor cortex to produce sounds. In such approach, cortical signals would be decoded into articulatory features to control in real time a parametric articulatory-based speech synthesizer having enough degrees of freedom to ensure continuous intelligible speech production. Interestingly, these articulatory features are generally considered lower dimensional and varying more slowly in time than acoustic features (Sondhi and Schroeter, 1987), thus possibly easier to predict from cortical signals.

In order to synthesize speech, an articulatory-based speech synthesizer that converts articulatory trajectories into an audible speech signal is needed. This topic is covered in the next chapter.

Chapter 2: Articulatory-based speech synthesis

As motivated in the previous chapter, our long term goal is too develop a brain-computer interface that decodes neural activity into control parameters for a speech synthesizer. We will consider neural activity from the speech motor cortex (see Chapter 1) because it was shown to be functionally organized by speech articulator – such as the lips or tongue – with activity correlated to articulatory rather than acoustic features during speech production (Bouchard et al., 2013; Cheung et al., 2016). While neural data could be decoded directly into acoustic parameters, we make the hypothesis that a relevant strategy could be to decode cortical signals to control in real time a parametric articulatory-based speech synthesizer, i.e. a synthesizer that produces speech not from text of phonetical input, but from information on the movements of the speech articulators. While there now exist high quality text-to-speech synthesis solutions, this was not the case for articulatory-based synthesis and was thus a first goal of my thesis. Before describing this work, I will review in this chapter the different types of speech synthesis that already exist, before focusing on articulatory-based speech synthesis and ways to acquire articulatory data needed for this type of synthesis.

I. Introduction

1. Speech production

The way speech is produced by our speech organs can be compared to the way notes are played using a wind instrument. The respiratory system plays the role of the bellows by expelling air through the trachea. At the upper part of the trachea is the larynx, our vibrator device, which is crossed by the vocal folds: during the production of an unvoiced sound, like /s/ or /v/, the vocal folds are opened and the air can freely go through; during the production of a voiced sound, like /a/ or /m/, they periodically open and close leading to the production of a periodic air wave. This modulated air flow then passes through the vocal tract, which plays the role of soundboard, and transforms the incoming air wave into speech. The vocal tract is composed by oral cavities, which geometry changes thanks to mobile organs, the vocal tract articulators. The main articulators are the tongue, the lips, the jaw and the soft palate (Fig. 6).

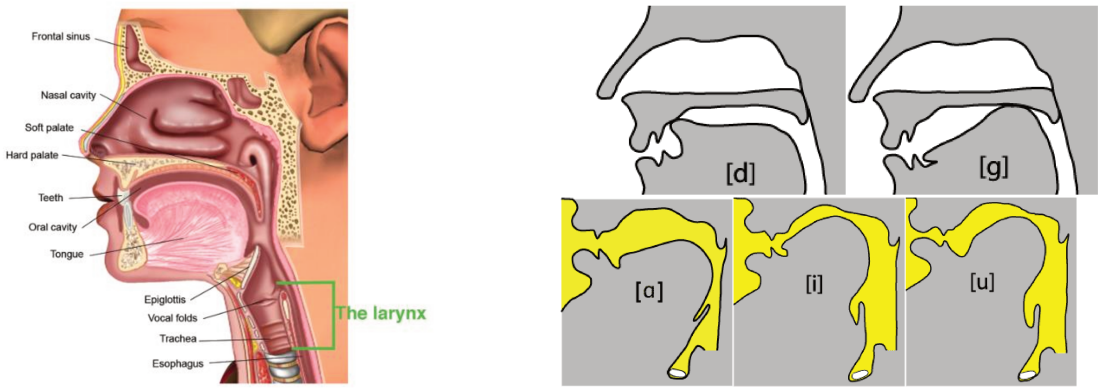


Fig. 6: Anatomy of the vocal tract and its configuration for different phonemes. Left – the main speech articulators are the tongue, the lips; the soft palate, the larynx and the teeth (Bluetree Publishing ©2013). Right – Different vocal tract configuration are shown, for /d/, /g/, /a/, /i/ and /u/ (www.indiana.edu).

Different configurations of these articulators lead to the production of different speech sounds (**Fig. 6**), and the phonemes are often categorized by their place of articulation (Kenyon, 1929).

2. Speech synthesis

Making machines that can speak has been a human dream for centuries, and the first attempts in order to produce artificial speech were made more than two hundred years ago, in 1769, when Wolfgang Ritter von Kempelen worked on mechanical models that could produce a variety of speech sounds (Schroeder, 1993). Thus, the first speech synthesis attempts were not made using electronics or computer science, but by carving pieces of wood to form resonators similar to the human vocal tract, like those made by Christian Kratzenstein in 1779 (Fig. 7).

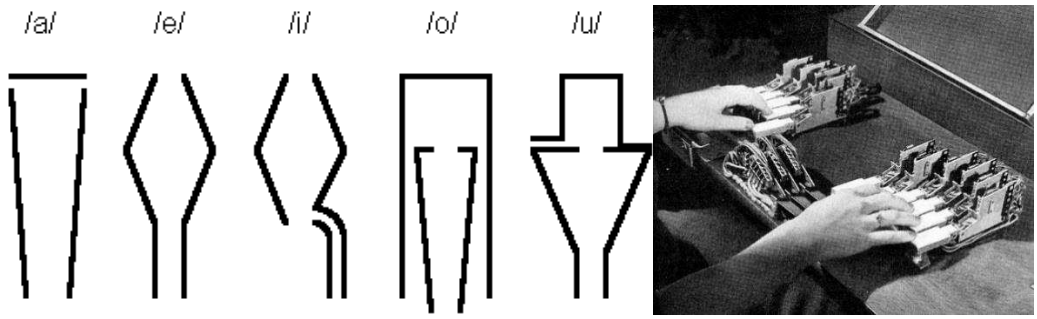


Fig. 7: Kratzenstein's resonators shapes and the Voder. Left – each vocal tract shape produces a different vowel when air is blown through them (Schroeder, 1993). Right – The Voder is controlled through a keyboard and some wrist controls.

In the 1930s, the Vocoder was developed in order to encode speech for better transmission in telecommunications systems (Dudley, 1939). The Vocoder was an electronic device that analyzed the changes of the speech signal spectral characteristics through time by passing it through a multiband filter in order to obtain an instantaneous representation of the spectral energy content. That way, a complex and fast varying speech waveform could be decomposed into values that changed slower over time, thus allowing to save bandwidth for transmission. This process could be reversed by filtering a broadband noise signal through filters depending on the encoded values. This Vocoder was later integrated to a demonstrator, the Voder, which was controlled by an operator using a keyboard (**Fig. 7**), and could produce some intelligible speech and even sing thanks to prosody and pitch control. Nowadays, a tool for processing speech signals is named a "vocoder", which can be used for speech synthesis, or for speech manipulation, or for speech encoding, etc.

The first computer-based speech synthesizers appeared in the 1950s (Schroeder, 1993), and have been used extensively since then, mainly for text-to-speech synthesis which converts written text into speech signal. The growing power and storage capacities of computers allowed for the use of more complex speech synthesis algorithms relying on large speech datasets.

Modern speech synthesis approaches can be achieved in several ways. One way to classify the different types of speech synthesis systems is by the type of their input parameters. We can thus distinguish three main categories of speech synthesis: formant synthesis, text-to-speech synthesis, and articulatory-based speech synthesis.

II. Formant synthesis

Formant synthesis uses input parameters that directly describe the spectral content of the target speech signal. In that case, the formant trajectories are explicitly specified as well as other features related to the glottal activity, for instance specifications about the unvoiced or voiced speech segments, or the fundamental frequency for voiced sounds. Most of the vocoders that are used for speech coding in telecommunication systems have a similar principle, except for the parametrization of the spectral content. The synthesis is done by modulating an excitation signal that represents the glottal activity through a time-varying filter, in order to obtain a speech waveform. That filter represents the transfer function of the vocal tract, the so-called spectral envelope. Different techniques can be used to model the spectral envelope, like linear predictive coding (Atal, 2006), mel-cepstral analysis (Imai et al., 1983), the "Harmonic+Noise" model (Laroche et al., 1993), or the "STRAIGHT" technique (Kawahara, 1997).

III. Text-to-speech synthesis

The most known category of speech synthesis systems is text-to-speech synthesis, which input is typically a sequence of words. It is generally made of two parts. First, a "natural language processing" module converts naturally written text into a sequence of phonetic units, like phonemes, along with some additional features describing its linguistic context (grammatical function if the unit is a word, or surrounding phonemes in the case of a phoneme, etc.) that are needed to determine the target prosody. The second module generates the speech waveform using the output from the first module (i.e. the segmentation of the input text into a sequence of phonetic units and the linguistic context features). In order to achieve that goal, different approaches exist that can be mostly divided into two large categories. First, concatenative synthesis (also called unit selection synthesis) consists in concatenating audio speech segments (such as diphone, demi-syllable, syllable, triphone or polyphone) selected from a large set of pre-recorded sentences. Second, statistical parametric synthesis (also called model-based synthesis) does not store any speech sample but stores a model that is used to convert input speech parameters – such as a sequence of phones, or some articulatory trajectories – into a speech signal, generally represented by parameters like the ones used by a vocoder.

1. Concatenative synthesis

Concatenative synthesis generally produces the most natural-sounding synthesized speech since it uses real speech samples. Main types of concatenative synthesis are diphone synthesis and unit selection synthesis.

Diphone synthesis makes use of a small database that contains one sample of each diphone that occurs in a given language, which number can range from several hundred to several thousand depending on the language. Prosody is added using signal processing techniques like linear predictive coding (Atal, 2006), PSOLA (Charpentier and Stella, 1986) or discrete cosine transform (Narasimha and Peterson, 1978). While diphone synthesis has the advantage to rely

on a very small database, it suffers from sonic glitches due to the concatenation of speech samples, and is rarely used nowadays.

On the other hand, unit selection synthesis relies on the use of large databases of speech samples in which each utterance is segmented into several different phonetic units like phones, diphones, syllables, words or sentences, and labeled with other phonetic properties – like pitch or duration – or even syntactic and lexical information. The phonetic segmentation can be done manually or automatically using forced alignment by automatic speech recognizer (Wagner, 1981; Brugnara et al., 1993) with further manual corrections. To perform the synthesis, the output speech signal is generated by determining the best chain of candidate units from the database using an index and search algorithms (Hunt and Black, 1996). When using a very large database, unit selection synthesis does not need to apply a lot of signal processing techniques which makes it sound very natural.

2. Statistical parametric synthesis

In statistical parametric synthesis, the spectral content of the speech signal is parametrized using one of the spectral envelop models mentioned before (c.f. “Formant synthesis”), and then the time-trajectories of the parameters are modeled for each phonetic class and linguistic context, generally using Hidden Markov Models (HMMs) (Tokuda et al., 1998), Deep Neural Networks (DNNs) (Ze et al., 2013) or similar models (Black et al., 2007). To perform the synthesis, the parameters of the spectral envelope model and the glottal parameters (like the fundamental frequency) are inferred and then used by the corresponding vocoder to generate the final speech waveform. Recently, artificial neural networks were even used to directly predict the speech waveform from text input, without passing by an intermediate representation of the spectral envelope (Oord et al., 2016).

Since this model-based approach does not use any speech sample at runtime, it can provide access to a wider obtainable sound-space with lower memory and processing requirements than concatenative speech synthesis, once the model is trained. However, a parametric representation of the speech signal is needed, and the reconstruction process, based on a vocoder, is often not ideal so that the resulting speech signal sounds less natural than when using a samples database.

IV. Articulatory-based speech synthesis

Another way to synthesize human speech is to directly simulate the physical principles of speech production, which is called “articulatory speech synthesis”. In that case, the input parameters are sequences of articulatory features, like the position of the main speech organs such as the tongue, the jaw, the lips, the velum or the larynx. Two main approaches have been proposed to synthesize speech from articulatory data: physical approaches that try to model the geometry of oral cavities and their acoustic properties, and the non-physical approaches that exploit large articulatory-acoustic databases and machine learning techniques to model the relationship between the articulatory movements and the corresponding speech signals. In both cases, a preliminary step consist in acquiring articulatory data, which can be achieved in several ways.

1. Methods for articulatory data acquisition

While there is a common agreement that the acoustic speech signal can be fairly recorded using a microphone, there is not a unique way to acquire articulatory data, and several methods have been proposed over the years in order to measure the vocal tract shape and its movements.

a. X-ray imaging

Traditional X-ray imaging can be used to acquire entire head images with good spatial (about 1-2 mm) and temporal resolution (about 50 frames/sec). This X-ray cineradiography was used for the first time in 1928 in order to study vowels productions (Russell, 1928), and has been originally the main source of information for analyzing the movement of the articulators during speech production. Since the whole vocal tract is visible, the entire shapes of the articulators can be extracted (Fig. 8).

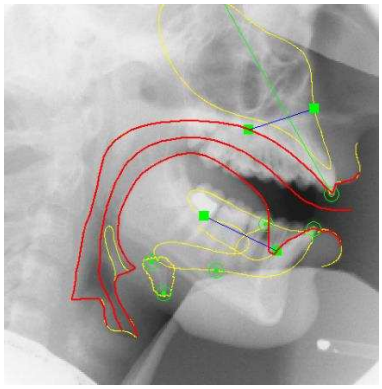


Fig. 8: Vocal tract shape extraction from X-ray cineradiography. The red, yellow and green marks show the manually extracted vocal tract shapes. Source: XArticulator software by Yves Laprie (https://members.loria.fr/YLaprie/ACS/index.htm).

However, identification of vocal tract structures in X-ray images is difficult since different head structures are projected on the same sagittal plane. This can be compensated by asking the subject to take contrast agents that adheres to the surface of the tongue, the mouth floor and the lips, and makes them easier to distinguish. Even so, it is still necessary to manually (Badin et al., 1995) or automatically (Thimm and Luettin, 1999; Fontecave Jallon and Berthommier, 2009; Laprie and Berger, 2015) segment the acquired images in order to extract the shape of the vocal tract articulators. Moreover, the exposure radiation time has to be limited for health safety reasons, which prevents recording large datasets of articulatory-acoustic data.

To extend the exposure time, an X-ray microbeam system was developed, which uses a narrow beam of X-ray to track the movements of small gold pellets attached to the speaker's articulators (Kiritani, 1986). That way, the exposition to radiation is reduced, which is safer for the subject and allows longer experiments, while still covering the whole vocal tract. Despite the reduced risk for the subject, this method has been largely replaced by safer methods such as magnetic resonance imaging (MRI) or ultrasonography.

b. Magnetic Resonance Imaging

Magnetic resonance imaging can be used to acquire the entire vocal tract shape in three dimensions, which allows direct calculation of vocal tract area and volume, and without any known dangerous effect for the subject. However, because of the way MRI scanners are constructed, the subjects have to be in supine position lying on their back, and articulatory movements are affected by the gravitational effects of this position (Stone et al., 2007). Moreover, to achieve high resolution (about 1mm) it is necessary for the subject to keep the same position for several seconds due to the very slow acquisition speed of MRI, which results in hyper-articulation (Engwall, 2003). Real-time MRI is now possible with frame rates as high as 500Hz (Uecker et al., 2010). It can even be achieved in three dimensions, but with lower spatial resolution (about 3mm) and lower frame rate (about 10 frames/sec) (Zhu et al., 2013). As for X-ray imaging, manual or automatic image segmentation is needed in order to obtain the shapes of vocal tract articulators.

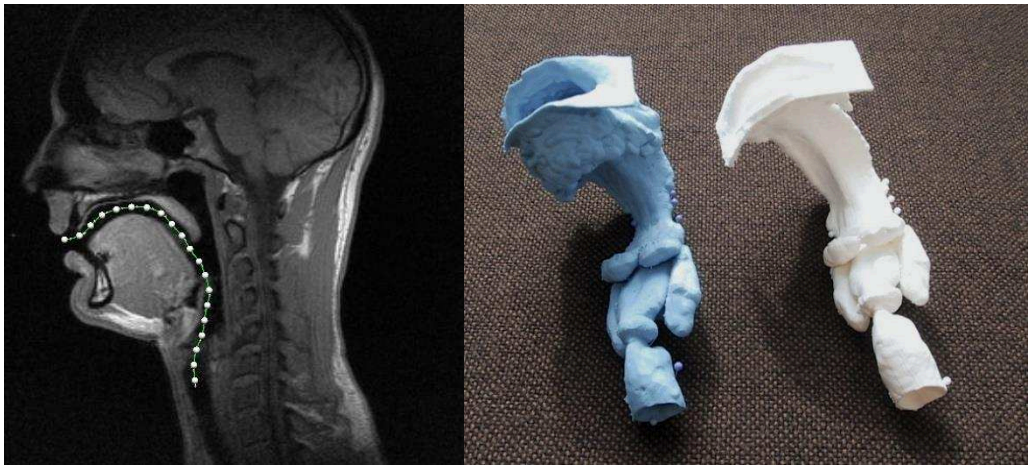


Fig. 9: vocal tract shape extraction from MRI. Left – image obtained by MRI, and automatically extracted shape (green line with white dots). Source: www.cmiss.bioeng.auckland.ac.nz. Right – 3D printed vocal tracts using three dimensional MRI data. Source: www.speech.math.aalto.fi

c. Video recording

One of the simplest way to acquire articulatory movements is to directly record a video of the subject while he is speaking. Thus, video recording of the lips has already been used for articulatory speech synthesis, mainly for vowel synthesis (Hasegawa and Ohtani, 1992). However, this technique only provide information about the external articulators, mainly the lips, which is not enough to discriminate all the phonemes of a language (Fisher, 1968), with state-of-the-art approaches generally reaching about 50-70% of word accuracy (Potamianos et al., 2003; Wand et al., 2016). That is why it is mostly used in combination with other techniques, such as ultrasonography (Hueber et al., 2010b).

d. Ultrasonography

Ultrasonography is an imaging technique that uses a high-frequency sound (ultrasound) wave, and estimates the delay between ultrasound pulses and their reflection to visualize internal body structures like muscles. It can be used through the lingual soft tissues in order to

visualize the articulatory movements of the tongue in real-time (Stone et al., 1988), and can be extended to three-dimensional acquisition (Deng et al., 2000; Fenster, 2001) without any known dangerous effect for the subject. The images obtained using ultrasonography are of poor quality but the surface of the tongue is mostly visible as a bright line on a black background (Fig. 10), so that its shape can be extracted manually or using automatic approaches (Fabre et al., 2015). As opposed to MRI, ultrasonography does not require the subject to be lying on its back. Nonetheless, the lack of visibility of tongue apex, the tongue walls contacts and multiple reflections make the automatic image processing difficult. Moreover, this technique only gives access to the tongue shape and is thus commonly combined with other modalities, such as video recordings, in order to obtain information from other articulators (Hueber et al., 2010b).

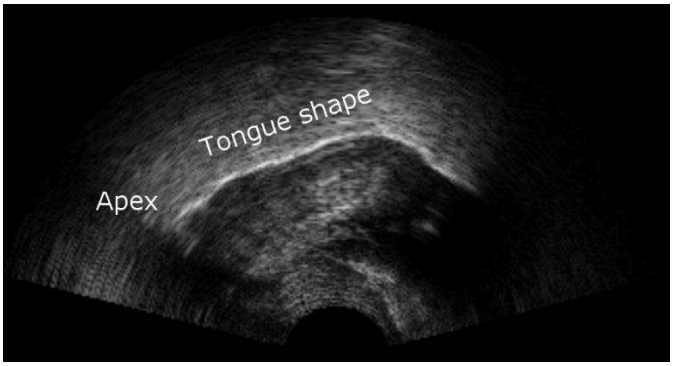


Fig. 10: Image of the tongue obtained by ultrasonography. The tongue surface is visible as a white line, but the tongue apex remains difficult to localize with precision (Hueber, 2009).

e. Electromyography

Electromyography (EMG) is a technique that enables to record the electrical activity produced by skeletal muscles using several electrodes (Fig. 11). Many muscles are implied in speech production, like the tongue and larynx muscles, and can be simultaneously recorded (Baer et al., 1988). EMG does not give direct access to the vocal tract articulators and movements, and mapping EMG signals to actual articulators position remains a difficult task: EMG signals do not represent the activation of a single muscle but a combination of various muscles, and the fibers of the same muscles are not activated at the same time, which make the final EMG signal a combination of various muscles and fibers signals (Jorgensen and Dusan, 2010). However, several studies showed that EMG signals contain enough information to be used for automatic speech recognition and direct speech reconstruction (Sugie and Tsunoda, 1985; Maier-Hein, 2005; Toth et al., 2009; Jorgensen and Dusan, 2010; Wand et al., 2013; Cler et al., 2014; Diener et al., 2016).

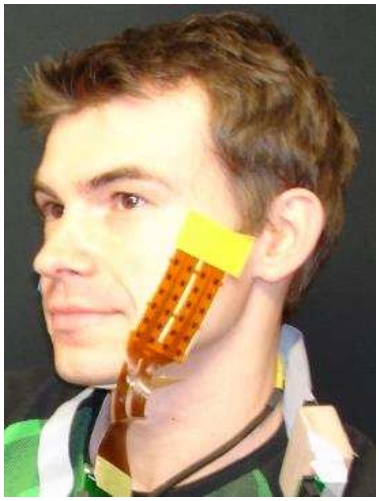


Fig. 11: Articulatory data from EMG. Here, an EMG array was placed to record cheek muscles activation (Wand et al., 2013).

f. Electropalatography

Electropalatography (EPG) can be used to monitor contact points between the tongue and the hard palate during speech production. This technique uses an artificial palate that is molded to fit against a specific speaker's hard palate, which exposes electrodes (from ten to hundreds) facing the tongue surface (Fig. 12). When the tongue contacts an electrode, an electric signal is transmitted which provides direct information in real-time on the tongue contact points with the palate which are crucial for the production of some consonants (Hardcastle and Roach, 1979). The need to mold a specific palate for each subject makes EPG difficult to use in numerous experiments – although 3D printing could solve this issue in the future. Also, while EPG provides information on contact points between the tongue and the hard palate, it does not provide any information on the global tongue shape and other articulators. However, EPG can be used in combination with other techniques of articulatory data acquisition, like electromagnetic articulography, in order to synthesize speech (Kello and Plaut, 2004).

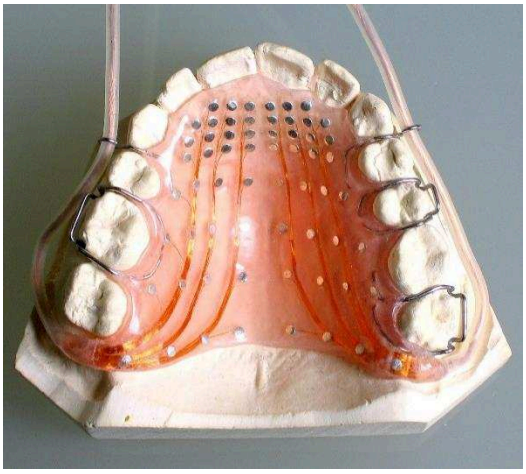


Fig. 12: Artificial palate for electropalatography. Each metal disk is an electrode that allows to detect contacts of the tongue with the palate. Source: www.articulateinstruments.com.

g. Electromagnetic articulography

Electromagnetic articulography (EMA) allows three-dimensional tracking of small sensor coils when placed near a magnetic field generator, with high spatial (< 1mm) and temporal (up to 400 frames/sec) resolutions. Several sensor coils can be glued on the tongue, the lips, the jaw and the soft palate in order to record the movements of the vocal tract articulators while the subject is speaking (Fig. 13). In EMA, several induction coils – different than the sensor coils – are placed near the head of the subject, and are supplied with current running at different frequencies for each coil. Each induction coil thus produces a variable electromagnetic field at a specific frequency which induces currents in the sensor coils that oscillate at the same frequency, and that is inversely proportional to the cube of the distance between the sensor coil and the induction coil. Thus, the composite induced current in each sensor can be separated out to determine the distance from each induction coil. Then, triangulation enables to determine its location in space. The EMA data can be retrieved in real-time for further processing, and a reference sensor allows to obtain all the coordinates with regards to an invariable reference point, even if the subject's head is moving.

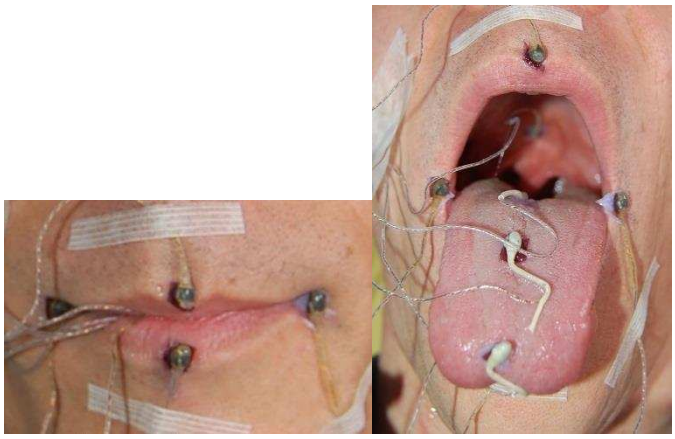


Fig. 13: Electromagnetic articulography. EMA coils are glued to the lips, the jaw, the tongue and the soft palate (Bocquelet et al., 2016a).

However, the EMA system comes with its own practical limitations. First, it is difficult to keep the coils fixed during long recording sessions, and the fixation duration depends on each subject, essentially because of salivation. Second, it is difficult – to not say impossible – to reattach the EMA coils at the exact same positions between two sessions, so that all the data has to be collected in a row. Third, some subjects have more sensitive soft palates than others, so that it can be impossible to place a coil on it in a comfortable way. Finally, the coils need electric wires to transmit data, which have to transit through the lips and might hinders articulation.

Nonetheless, EMA allows direct access to the positions and movements of the articulators with high precision and very good time resolution, which makes it a very valuable tool for analyzing speech articulatory movements. Because of its low invasiveness, it has very low risks for the subject with regards to X-ray imaging. Moreover, not only the position of the sensor coils is available, but their orientation as well, which can provide additional information about the vocal tract shape.

There exists another similar method based on electromagnetics in order to track movements of the vocal tract articulators, which is referred to as "permanent-magnetic articulography" or "Tongue Drive" (Huo et al., 2008). In this approach, a small cylindrical permanent magnet has to be secured on the tongue by implantation, piercing or tissue adhesives, and a pair of three-axial linear magneto-inductive sensor modules has to be mounted bilaterally on a headset near the subject's cheeks. This sensor wirelessly transmits the magnetic field information to a computer which uses an electromagnetic model in order to predict the magnet position and orientation. This system allows real time tracking (about 20 frames/sec) with good spatial resolution (about 1mm) (Cheng et al., 2009). Although it has the advantage of being wireless with regards to EMA, it is limited to one sensor, which is clearly not enough in order to capture the whole vocal tract shape. Extensions with several magnets exist but do not allow to revert the recorded electromagnetic signals into the sensors positions (Fagan et al., 2008).

h. Choice of an articulatory data recording method

In order to choose which articulatory data acquisition method is more suitable for articulatory speech synthesis, it is necessary to analyze the acoustic and articulatory properties of natural speech. Phonemes duration in normal speech typically range from about 10ms for plosive consonants to more than 100ms for vowels (Kuwabara, 1996; de Mareüil et al., 2008; Ziolko and Ziolko, 2011). In order to capture all the dynamics of natural speech production it is thus necessary to have a recording system with a good temporal resolution, at least superior to 100Hz. Moreover, speech production requires precise gestures and a displacement of the articulators by a few millimeters can result in producing a totally different sound (Perrier, 2005). Therefore, sufficient spatial resolution is needed in order to capture discriminable position of the articulators while producing distinct phonemes. Since the tongue, jaw, lips and soft palate are all involved differently in speech production, it is necessary to be able to record the whole vocal tract shape or information that can represent these articulators shape. Finally, 3D data acquisition might be desirable especially for lateral consonants.

The comparison of the previously mentioned acquisition procedures is summarized in **Table 2**. X-Ray imaging presents risks for the subjects which make it usable only for a short period of time; MRI has a moderate spatial resolution but requires manual or semi-automatic segmentation of images and the subject has to be in supine position; ultrasonography and EPG only give information about the tongue and have to be combined with other methods in order to get the full vocal tract shape; video recording only give information about the external articulators such as the lips; and EMG does not give direct access to articulators positions or shape but only to muscles activations. Thus, and regarding the previous considerations, the choice was made to use electromagnetic articulography for acquiring the articulatory data, since it provides good spatial and temporal resolution, even if it is a slightly invasive technique (gluing of sensors on the speech articulators) and does not allow access to the full vocal tract shape but only to the 3D position of several points on the speech articulators.

	X-Ray	MRI	Ultra-sonography	EPG	EMG	Video	EMA
Indicative time resolution	30-60 Hz	0-200 Hz	200 Hz	200 Hz	5-500 Hz	24-10,000 Hz	100-400 Hz
Indicative spatial resolution	1-2 mm	1-3mm	-	3 mm	-	1 mm	< 1mm
3D data	No	Yes	Depends	No	-	No	Yes
Tongue imaging	Profile shape	Full shape	Profile shape	Contact points with the hard palate	Muscles activation	None	Several points at the surface
Lips imaging	Profile shape	Full shape	None	None	Muscles activation	Front view	Several points
Jaw imaging	Yes	Yes	None	None	Muscles activation	None	Yes
Velum imaging	Yes	Yes	None	None	None	None	Yes
Risk / Invasiveness	High	None	Very low	Low	Very low	None	Low
Cost	High	High	Low	Low	Low	Low	Low

Table 2: Comparison of different articulatory data acquisition methods. Completed from (Youssef, 2011). Colors are indicative and reflect the compliance of the method with regard to each criteria.

2. Physical modeling of the vocal tract

A first approach to synthesize speech from articulatory data is to realistically mimic the functioning of our speech organs. Since speech can be seen as the result of the modulation of an excitation air wave by variable geometry pipe, acoustic simulation and modeling technique could be used to artificially synthesize speech, which first requires to model the geometry of the oral cavities and then their acoustic properties.

a. Modeling the geometry of oral cavities

The vocal tract is made of several cavities which geometry changes due to the movement of articulators. In order to simulate the acoustic properties of the vocal tract and synthesize speech, geometric models of the vocal tract controllable by few parameters are needed. The vocal tract models can be categorized into two-dimensional (2D) and three-dimensional (3D) models on one hand, and into geometrical, statistical, and biomechanical models on the other hand.

i. 2D and 3D models of oral cavities

The 2D models define the vocal tract geometry in the midsagittal plane, using the contour lines of the articulators (Joseph S. Perkell, 1974; Maeda, 1990; Payan and Perrier, 1997). However, such models do not provide any information about the cross-sectional shape of the vocal tract which varies greatly along its length. In fact, the shape itself has almost a negligible effect on the resonance properties of the vocal tract, the most important being the changes in cross-sectional area of the vocal tract along its length. In most cases, this area is estimated by making the approximation that it is circular or elliptic (Maeda, 1990).

The 3D models do not suffer from this approximation since they give access to whole 3D geometry of the vocal tract (Engwall, 1999; Dang and Honda, 2004; Birkholz et al., 2006; Perrier et al., 2011). Moreover, they can represent configurations that cannot be represented by a 2D model, like lateral consonants as /l/, for which the air flows along the sides of the tongue, while it is blocked by the tongue in the middle of the mouth.

ii. Geometrical, statistical and biomechanical models of oral cavities

Geometric models directly define the vocal tract shape using few geometrical parameters that were chosen a priori – for instance the aperture of the mouth or the height of the tongue tip, and can be fitted a posteriori to particular data (Engwall, 1999; Birkholz et al., 2006).

Statistical models use large amount of vocal tracts shapes, generally obtained from MRI data, to extract uncorrelated parameters that can represent the geometry of the vocal tract, using statistical techniques such as Principal Component Analysis (Maeda, 1990). A statistical model is often dedicated to the particular speaker from whom the data is extracted.

Finally, the biomechanical models aim to simulate the behavior of the vocal tract articulators and muscles by using finite elements methods (Dang and Honda, 2004; Perrier et al., 2011). Their main purpose is to study the relationship between muscle activation and the articulatory movements, and the finite element approach requires a significant computational power and has a high number of degrees of freedom (DoF).

Once the vocal tract geometry is known, numerical simulations of wave propagation through the vocal tract shape can be used to generate a speech signal from an articulatory configuration. This requires to model the acoustic properties of the oral cavities.

b. Modeling the acoustic properties of oral cavities

Different methods have been proposed to model the acoustic properties of the oral cavities in order to generate speech from the geometry of the vocal tract.

Most approaches model the vocal tract as a series of circular sections which area is obtained using an appropriate model of the vocal tract geometry (c.f. previous section). In that case, the vocal tract is discretized along its length so that its shape can be approximated by concatenating circular tubes of the same length. The nasal cavity is often considered as an on/off air path: if the air passes through the nasal cavity, an additional area is added to the section of the vocal tract after the nasal branching (Maeda, 1982). An acoustic model is then applied in order to

simulate the resonance contributions of all the sections, and the acoustic equations are solved either in the frequency-domain, or in the time-domain or using hybrid approaches.

In the frequency-domain methods, the acoustic transfer function of the modeled vocal tract can be obtained using the area function, i.e. the evolution of the vocal tract area along its length (Fant, 1975; Rubin and Baer, 1981; Ngoc and Badin, 1994). This calculation is mostly based on the Kelly-Lochbaum model (Kelly and Lochbaum, 1962) which models the wave propagation and takes into account frequency-independent propagation losses within sections and reflections at the section boundaries. The speech signal can then be obtained using the source-filter approach: the speech signal spectrum can be expressed as the multiplication of the source spectrum by the vocal tract transfer function and other acoustic transfer function, for instance to take into account the lip radiation effect. This approach is particularly used for stationary sounds, but can be extended to dynamic geometries (Nowakowska et al., 1993). In (Hasegawa and Ohtani, 1992), this approach was used to synthesize five Japanese vowels with a recognition rate of about 91% from video recordings of the lips: the shape of the lips was extracted from the video images and then used to estimate the area function, which was in turn used to synthesize speech.

In the time-domain methods, the output waveform is obtained by directly applying a set of equations to an input excitation signal across time (Flanagan et al., 1975; Maeda, 1982). The fact that the equations are directly solved in the time domain makes it directly applicable for synthesizing speech while changing the vocal tract geometry over time.

Hybrid methods try to combine advantages of both time-domain and frequency-domain methods. For instance, a frequency-domain method can be applied to estimate the vocal tract transfer function while the glottal excitation is obtained in the time-domain (Sondhi and Schroeter, 1987).

Most of the previously mentioned acoustic simulations are one-dimensional (1D), assuming plane wave propagation only. However, three-dimensional acoustic simulating methods can be used in order to obtain more precise characteristics of the vocal tract by taking advantage of 3D geometry models (Kagawa et al., 1992; Takemoto et al., 2010; Svec et al., 2011). Some studies pointed out that 3D simulation methods could exhibit resonance modes that are not observable using 1D simulation (Takemoto et al., 2014). Nonetheless, while 1D methods can generally be applied in real-time, 3D methods often rely on heavy computation such as finite elements approaches, which make them harder to use in real-time applications. This is a potential issue for future use in a brain-computer interface for speech rehabilitation. Moreover, speech is a very complex phenomenon while physics approaches need to work under several assumptions and approximations in order to simplify models. By contrast, machine learning-based approaches do not model the physical and acoustic properties of the vocal tract.

3. Non-physical articulatory-based synthesis

While physical approaches try to model the biophysics behind the speech production, the non-physical approaches usually use supervised machine-learning methods in order to model or capture the relationship between articulatory and acoustic observations (so called “articulatory-to-acoustic mapping”), by exploiting large databases of synchronously recorded articulatory and acoustic data. These approaches can be mainly divided into two categories:

those that first use articulatory speech recognition, i.e. that recognize sequences of phonemes or words from the articulatory data, which are then used to synthesize speech for instance using text-to-speech synthesis methods, and those that "directly" estimate the acoustic trajectories – i.e. the time-varying sequences of acoustic parameters – from the articulatory trajectories using statistical inference that can then be converted into a speech waveform using the corresponding vocoder.

a. Discrete approaches

The first type of approaches uses an intermediate representation – such as a sequence of phones or words – which is then used to synthesize the final speech signal, for instance using text-to-speech synthesis. Such approach was originally performed on electromyography (EMG) data from three sensors placed on the speaker's face, allowing to recognize five isolated Japanese vowels with 71% accuracy (Sugie and Tsunoda, 1985). In (Jorgensen et al., 2003), EMG from four electrodes over the throat and electropalatography (EPG) data were used to classify six isolated words with about 90% accuracy, using different signal representations and classification algorithms, such as artificial neural networks (ANNs) or linear classifiers. In (Maier-Hein et al., 2005; Jou et al., 2006, 2007; Walliczek et al., 2006), twelve EMG electrodes were placed over the face, throat and chin of the speaker to record the muscle activity of the lips and tongue muscles in order to decode continuous speech into sequences of phones. The use of an intermediate phonetic representation allows to take advantage of linguistic knowledge to constrain the recognition to meaningful sequences of phonetic units. For instance, a limited vocabulary can be used to force the recognition to a set of predefined words, thus increasing the accuracy of the synthesis, or grammar rules can be used to generate only grammatically correct sentences, further increasing the accuracy of the system. In this study, the authors constrained the vocabulary to a set of 100 words and used a trigram language model, allowing to reach a word recognition accuracy of about 70%. More recently permanent-magnetic articulography was used to recognize words in a finite vocabulary of 9 words, or phones among a limited set of 13 phones (Fagan et al., 2008). Seven permanent magnets were placed on the tongue, lips and jaw of a speaker while six magnetic sensors mounted on glasses recorded changes in the magnetic field produced by the magnets movements. Words or phones were recognized by comparing them with known data, allowing to reach 97% accuracy for words and 94% for vowels. However, the size of the vocabulary for words was limited to 9 words only, and the phones were isolated, which means that lower accuracy has to be expected in the case of full words or sentences.

b. Continuous approaches

The second type of approaches generally considers statistical models to perform the articulatory-to-acoustic mapping, such as artificial neural networks (ANNs), Gaussian mixture models (GMMs) or hidden Markov models (HMMs). The fundamental basis of GMMs and ANNs are presented in sections V and VI of this chapter.

In (Kello and Plaut, 2004), the authors presented a single hidden layer ANN model that was trained on the MOCHA database to directly predict the power spectrum of the speech signal. The MOCHA database contained a combination of electromagnetic articulography

(EMA), electropalatography (EPG) and laryngography (recording of the glottal activity) recordings from one British English speaker, for a total of about 460 different sentences. The synthesis intelligibility was assessed by listeners and achieved a word recognition rate of about 60% when tested on data that was not part of the training data used to compute the neural network parameters. While being very promising, and considering that there were not isolated words but rather words within meaningful sentences, such synthesis could not be considered as fully intelligible. In (Denby and Stone, 2004), ANNs were used on tongue contours extracted from ultrasonography images to control a GSM (Global System for Mobile communications) vocoder. While similarities between original and synthesized sounds could be observed, this did not lead to intelligible synthesis. The authors later suggested to combine ultrasonography with video recordings of the lips in order to improve the synthesis intelligibility (Denby et al., 2006). This was done in a thesis work (Hueber, 2009) where the ANN-based mapping is compared with a GMM-based mapping. This approach was originally proposed by Toda et al. to perform the articulatory-to-acoustic mapping as well as the inverse problem of the acoustic-to-articulatory mapping using only the EMA and audio part of the MOCHA database (Toda et al., 2008). In particular, they proposed a way to take into account the dynamic properties of the target signal – here the acoustic trajectories – which they called the "trajectory GMM" (see section V). In this method, the determination of a target signal having appropriate static and dynamic properties is obtained by explicitly imposing the relationship between static and dynamic features. In this study, the intelligibility of the synthesis was not directly assessed through a transcription or a recognition task. However, synthesis examples provided with the paper showed that intelligible synthesis could be achieved.

Hidden Markov models (HMMs) have been used to perform the articulatory-to-acoustic mapping using articulatory data obtained from ultrasonography and video recording (Hueber et al., 2012; Hueber and Bailly, 2016). In these studies, a hidden Markov model was used to infer the most probable sequence of phones corresponding to input articulatory data, similarly to the first category of approaches. However, here the decoded sequence of phones was then combined with the input articulatory data to finally infer acoustic trajectories in order to synthesize speech. This allowed to take advantage of linguistic knowledge while still taking into account articulatory features when performing the synthesis.

EMA data and HMMs have been used in several studies for the inverse problem of the acoustic-to-articulatory mapping (Hiroya and Honda, 2004; Zen et al., 2011). In (Ling et al., 2009), articulatory features obtained from EMA data were integrated into a HMM-based text-to-speech synthesizer to improve or alter speech synthesis. However to our knowledge there are no studies considering HMM-based speech synthesis only from EMA data. On the other hand, GMMs and ANNs have already shown promising results on synthesizing speech from EMA data. Moreover, recent developments in the field of artificial neural networks, particularly with the emergence of deep neural network (DNNs), suggested that this approach might be improved to reach fully intelligible synthesis. In particular, DNNs were successfully used for the inverse problem of the acoustic-to-articulatory mapping, i.e. recovering articulatory trajectories from the speech signal (Uria et al., 2011).

In the present work, we thus chose the trajectory GMM approach as a gold-standard (Toda et al., 2008), and we considered DNNs for the articulatory-to-acoustic mapping. Both these approaches are presented in details in the following sections.

V. GMM-based articulatory-to-acoustic mapping

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. It is mainly used to estimate and model the probability distribution of continuous measurements, such as articulatory or acoustic data.

1. The trajectory GMM

The trajectory GMM approach was introduced by Toda et al. (Toda et al., 2008) for articulatory-to-acoustic mapping. In the training stage, the joint probability density function (pdf) of articulatory and acoustic data is modelled by a GMM, which is then used in the mapping stage to infer the more probable sequence of acoustic parameters for a given input sequence of new articulatory data.

a. Probability density function

More formally, if we note $\underline{\mathbf{x}} = [\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T]$ a sequence of T articulatory features vectors $\mathbf{x}_t$, and $\underline{\mathbf{y}} = [\mathbf{y}_1, \ldots, \mathbf{y}_t, \ldots, \mathbf{y}_T]$ a sequence of T acoustic features vector $\mathbf{y}_t$, then in the training stage the joint probability density $\underline{\mathbf{Z}}$ of the training articulatory data $\underline{\mathbf{X}}$ and acoustic data $\underline{\mathbf{Y}}$ is modelled by a mixture of M Gaussians of parameters $\boldsymbol{\Theta} = \{\alpha_1, \ldots, \alpha_M, \boldsymbol{\mu}_1^{\boldsymbol{z}}, \ldots, \boldsymbol{\mu}_M^{\boldsymbol{z}}, \underline{\boldsymbol{\Sigma}_1^{\boldsymbol{z}}}, \ldots \underline{\boldsymbol{\Sigma}_M^{\boldsymbol{z}}}\}$:

$$p(\underline{\boldsymbol{Z}}|\boldsymbol{\Theta}) = p(\underline{\boldsymbol{X}}, \underline{\boldsymbol{Y}}|\boldsymbol{\Theta}) = \sum_{m=1}^{M} \alpha_m N(\boldsymbol{z}, \boldsymbol{\mu}_{\boldsymbol{m}}^{\boldsymbol{z}}, \underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{z}}}) \quad \text{Eq. 1}$$

With $N(., \boldsymbol{\mu}, \underline{\boldsymbol{\Sigma}})$ a normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\underline{\boldsymbol{\Sigma}}$, and α_m the weight associated to the m[th] mixture component so that $\sum_{m=1}^{M} \alpha_m = 1$ and $\alpha_m \geq 0$.

The mean vector $\boldsymbol{\mu}_{\boldsymbol{m}}^{\boldsymbol{z}}$ and the covariance matrix $\underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{z}}}$ of the m[th] component can be written as:

$$\boldsymbol{\mu}_{\boldsymbol{m}}^{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{m}}^{\boldsymbol{x}} \\ \boldsymbol{\mu}_{\boldsymbol{m}}^{\boldsymbol{y}} \end{bmatrix} \quad \text{Eq. 2}$$

$$\text{And } \underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{z}}} = \begin{bmatrix} \underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{xx}}} & \underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{xy}}} \\ \underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{yx}}} & \underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{yy}}} \end{bmatrix} \quad \text{Eq. 3}$$

Where $\boldsymbol{\mu}_{\boldsymbol{m}}^{\boldsymbol{x}}$ and $\boldsymbol{\mu}_{\boldsymbol{m}}^{\boldsymbol{y}}$ are the mean vectors of the m[th] mixture component for $\underline{\mathbf{X}}$ and for $\underline{\mathbf{Y}}$, respectively. $\underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{xx}}}$ and $\underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{yy}}}$ are the covariance matrix of the m[th] mixture component for $\underline{\mathbf{X}}$ and for $\underline{\mathbf{Y}}$, respectively, and $\underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{xy}}}$ and $\underline{\boldsymbol{\Sigma}_{\boldsymbol{m}}^{\boldsymbol{yx}}}$ are the cross-covariance matrices of $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$. These parameters are estimated from training data, generally using the expectation-maximization (EM) algorithm.

b. Mapping function

In the mapping stage, for a given input sequence of new articulatory data $\underline{\mathbf{x}} = [\underline{\mathbf{x}_1}, \ldots, \underline{\mathbf{x}_t}, \ldots, \underline{\mathbf{x}_T}]$, the conditional probability density function (pdf) of each frame $p(\boldsymbol{y_t}|\boldsymbol{x_t}, \boldsymbol{\Theta})$ can be directly derived from the previously computed GMM:

$$p(\boldsymbol{y_t}|\boldsymbol{x_t}, \boldsymbol{\Theta}) = \sum_{m=1}^{M} p(\mathrm{m}|\mathbf{x_t}, \boldsymbol{\Theta}) \cdot p(\boldsymbol{y_t}|\boldsymbol{x_t}, m, \boldsymbol{\Theta}) \quad \text{Eq. 4}$$

$$\text{where } p(\mathrm{m}|\mathbf{x_t}, \boldsymbol{\Theta}) = \frac{\alpha_m N(\boldsymbol{x_t}, \boldsymbol{\mu_m^x}, \underline{\boldsymbol{\Sigma_m^x}})}{\sum_{i=1}^{M} \alpha_i N(\boldsymbol{x_t}, \boldsymbol{\mu_i^x}, \underline{\boldsymbol{\Sigma_i^x}})} \quad \text{Eq. 5}$$

$$\text{and } p(\boldsymbol{y_t}|\boldsymbol{x_t}, m, \boldsymbol{\Theta}) = N(\mathbf{y_t}, \boldsymbol{E_{m,t}^y}, \underline{\boldsymbol{D_m^y}}) \quad \text{Eq. 6}$$

with the mean vector $\boldsymbol{E_{m,t}^y}$ and the covariance matrix $\underline{\boldsymbol{D_m^y}}$ of the conditional distribution equal to:

$$\boldsymbol{E_{m,t}^y} = \boldsymbol{\mu_m^y} + \underline{\boldsymbol{\Sigma_m^{yx}}} \left(\underline{\boldsymbol{\Sigma_m^{xx}}}\right)^{-1} (\boldsymbol{x_t} - \boldsymbol{\mu_m^x}) \quad \text{Eq. 7}$$

$$\text{and } \underline{\boldsymbol{D_m^y}} = \underline{\boldsymbol{\Sigma_m^{yy}}} - \underline{\boldsymbol{\Sigma_m^{yx}}} \left(\underline{\boldsymbol{\Sigma_m^{xx}}}\right)^{-1} \underline{\boldsymbol{\Sigma_m^{xy}}} \quad \text{Eq. 8}$$

This conditional pdf can then be used to infer the best corresponding sequence of acoustic features vectors $\underline{\hat{\mathbf{y}}} = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_t, \ldots, \hat{\mathbf{y}}_T]$, given some criterion.

When using the Minimum Mean-Square Error (MMSE) criterion (Stylianou et al., 1998), the sequence $\underline{\hat{\mathbf{y}}}$ is estimated on a frame-by-frame basis as:

$$\hat{\boldsymbol{y}}_t = Expectation[\hat{\boldsymbol{y}}_t|\boldsymbol{x_t}] = \sum_{m=1}^{M} p(\mathrm{m}|\mathbf{x_t}, \boldsymbol{\Theta}) \cdot \boldsymbol{E_{m,t}^y} \quad \text{Eq. 9}$$

Hence, the predicted frame t of the sequence $\underline{\hat{\mathbf{y}}}$ can be directly computed as a linear combination of the mean vectors of the conditional pdf mean vectors (Eq. 9), weighted by the posterior probabilities that the vector $\boldsymbol{x_t}$ belongs to each one of the mixture components (Eq. 5).

However, (Toda et al., 2008) have shown that this inference using the MMSE criterion is not appropriate for multiple distributions since it does not take into account the covariance matrices of the conditional probability distributions shown in (Eq. 8), although they might be informative. Moreover, since the mapping is done on a frame-by-frame basis, it does not take into account the dynamic properties of the acoustic trajectories to be predicted. They thus proposed the trajectory GMM approach, using the Maximum Likelihood Estimation (MLE) criterion. When using the MLE criterion, the sequence $\underline{\hat{\mathbf{y}}}$ is estimated by:

$$\hat{\boldsymbol{y}}_t = arg \max_{y_t} p(\boldsymbol{y_t}|\boldsymbol{x_t}, \boldsymbol{\Theta}) \quad \text{Eq. 10}$$

Note that as opposed to when using the MMSE criterion, there is no closed form solution to directly compute $\hat{\boldsymbol{y}}_t$. However, this definition of $\hat{\boldsymbol{y}}_t$ now takes into account the covariance

matrices of the conditional probability distributions (see Eq. 10 combined with Eq. 4, Eq. 6 and Eq. 8).

In order to take into account the dynamic properties of the acoustic parameters $\underline{\mathbf{Y}}$, (Toda et al., 2008) introduce its N derivatives to define a new variable $\underline{\mathbf{W}}$ such that $\underline{\mathbf{w}} = [\mathbf{w_1}, \ldots, \mathbf{w_t}, \ldots, \mathbf{w_T}]$ with $\mathbf{w_t} = [\mathbf{y_t}, \Delta^{(1)}\mathbf{y_t}, \Delta^{(2)}\mathbf{y_t}, \ldots, \Delta^{(n)}\mathbf{y_t}, \ldots, \Delta^{(N)}\mathbf{y_t}]$ and $\Delta^{(n)}\mathbf{y_t}$ is the nth derivative of $\mathbf{y_t}$ using finite differences. For instance:

$$\Delta^{(1)}\boldsymbol{y_t} = \frac{\boldsymbol{y_{t+1}} - \boldsymbol{y_{t-1}}}{2} \quad \text{Eq. 11}$$

$$\Delta^{(2)}\boldsymbol{y_t} = \frac{\boldsymbol{y_{t+1}} - 2 * \boldsymbol{y_t} + \boldsymbol{y_{t-1}}}{4} \quad \text{Eq. 12}$$

This allows to define a matrix $\underline{\mathbf{A}}$ so that $\underline{\mathbf{W}} = \underline{\mathbf{A}}.\underline{\mathbf{Y}}$. Thus by replacing $\underline{\mathbf{Y}}$ by $\underline{\mathbf{A}}.\underline{\mathbf{Y}}$ in the previous equations, one can take into account the dynamic properties of the acoustic parameters when inferring the sequence $\underline{\hat{\mathbf{y}}}$.

For the mapping phase (Toda et al., 2008), a conditional pdf $p(\boldsymbol{Y_t}|\boldsymbol{x_t}, \boldsymbol{\Theta})$ is derived, for each frame t, from the joint pdf $p(\boldsymbol{x_t}, \boldsymbol{Y_t})$ estimated during training by a GMM, such as:

$$p(\boldsymbol{Y_t}|\boldsymbol{x_t}, \hat{m}_t, \boldsymbol{\Theta}) = N(\boldsymbol{Y_t}, \boldsymbol{E}^{\boldsymbol{y}}_{\hat{m}_t,t}, \underline{\boldsymbol{D}}^{\boldsymbol{y}}_{\hat{m}_t}) \quad \text{Eq. 13}$$

$$\text{with} \begin{cases} \boldsymbol{E}^{\boldsymbol{y}}_{\hat{m}_t,t} = \boldsymbol{\mu}^{\boldsymbol{y}}_{\hat{m}_t} + \underline{\boldsymbol{\Sigma}}^{\boldsymbol{yx}}_{\hat{m}_t} \left(\underline{\boldsymbol{\Sigma}}^{\boldsymbol{xx}}_{\hat{m}_t}\right)^{-1} (\boldsymbol{x_t} - \boldsymbol{\mu}^{\boldsymbol{x}}_{\hat{m}_t}) \\ \underline{\boldsymbol{D}}^{\boldsymbol{y}}_{\hat{m}_t} = \underline{\boldsymbol{\Sigma}}^{\boldsymbol{yy}}_{\hat{m}_t} - \underline{\boldsymbol{\Sigma}}^{\boldsymbol{yx}}_{\hat{m}_t} \left(\underline{\boldsymbol{\Sigma}}^{\boldsymbol{xx}}_{\hat{m}_t}\right)^{-1} \underline{\boldsymbol{\Sigma}}^{\boldsymbol{xy}}_{\hat{m}_t} \end{cases} \quad \text{Eq. 14}$$

Where $\hat{m} = [\hat{m}_1, \ldots, \hat{m}_t, \ldots, \hat{m}_T]$ is the suboptimum sequence of mixture components defined as $\hat{m} = \arg\max_m \{P(m|\boldsymbol{x}, \boldsymbol{\Theta})\}$ determined in our implementation using the Viterbi algorithm (Bishop and Christopher M. B, 2006). Finally, the output trajectories $\hat{\boldsymbol{y}}$ are estimated using the following equation:

$$\underline{\hat{\boldsymbol{y}}} = (\underline{\boldsymbol{W}}^T \underline{\boldsymbol{D}}_{\hat{m}}{}^{-1} \underline{\boldsymbol{W}})^{-1} \underline{\boldsymbol{W}}^T \underline{\boldsymbol{D}}_{\hat{m}}{}^{-1} \underline{\boldsymbol{E}}_{\hat{m}} \quad \text{Eq. 15}$$

with $\underline{\boldsymbol{E}}_{\hat{m}} = [\boldsymbol{E}_{\hat{m}_1}, \ldots, \boldsymbol{E}_{\hat{m}_T}]$ and $\underline{\boldsymbol{D}}_{\hat{m}}{}^{-1} = diag[\underline{\boldsymbol{D}}_{\hat{m}_1}{}^{-1}, \ldots, \underline{\boldsymbol{D}}_{\hat{m}_T}{}^{-1}]$.

2. Training algorithm for the trajectory GMM

In practice, during the training phase, the parameters of the GMM are estimated using the EM algorithm with the training dataset. The EM algorithm is an iterative method which needs an initial vector of parameters to optimize, the final optimization result depending on the quality of this first approximation. Here, the initial parameters were computed using the clustering k-means algorithm on the joint articulatory-acoustic data: this is an iterative and simple algorithm to partition data into k clusters in which each observation belongs to the cluster with the nearest mean. For a GMM with M Gaussian components, the joint articulatory-acoustic data $\underline{\mathbf{Z}}$ is first divided into M clusters. Here we run the k-means algorithm twice and kept the one with the minimum summed distance value. Then, the mean $\boldsymbol{\mu}^{\boldsymbol{z}}_{\boldsymbol{m}}$ and covariance matrix $\underline{\boldsymbol{\Sigma}}^{\boldsymbol{z}}_{\boldsymbol{m}}$ of the mth

cluster are used as initialization parameters for the m^{th} component of the GMM. The weight α_m of the m^{th} component is computed as the proportion of data points belonging to the m^{th} cluster.

VI. DNN-based articulatory-to-acoustic mapping

Deep Neural Networks (DNNs) are part of a larger category of mathematical models called Artificial Neural Networks (ANNs).

1. Artificial Neural Network

In its simpler form, an ANN is a mathematical model, originally inspired by the behavior of biological neurons, made up of computational units (or "neurons"). Each unit has one output and can have several inputs coming from the output of other units in the network, each one with an associated weight. The unit itself is defined by an activation function σ and a bias, so that the output of the unit is the result of applying the activation function σ to the weighted sum of its inputs, plus the bias (**Fig. 14**).

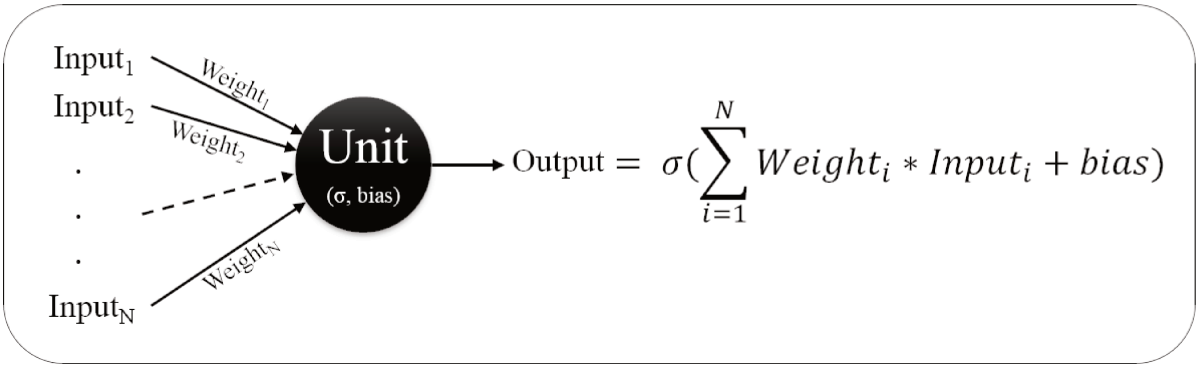


Fig. 14: Conventional unit of an artificial neural network. The unit has one output, and several inputs $Input_1$, ..., $Input_N$ each one with an associated weight $Weight_1$, ..., $Weight_N$. *A unit is defined by its activation function σ and its bias, so that the output* of the unit is *the application of σ to the weighted sum of its inputs, to which is added the bias.*

Units in an ANN are often organized into layers, so that each unit of a layer is connected to all units of the next layer, and there are no connections between units belonging to the same layer, forming a so-called feed-forward neural network (**Fig. 15**). More formally, the first layer is a visible input layer h_0, while the next L layers are hidden layers h_1, h_2, ..., h_L, and the last one is the output layer h_{L+1}. In the following, $w_{i,j}^{(l)}$ denotes the weight of a connection from unit $u_i^{(l-1)}$ of layer h_{l-1} to unit $u_j^{(l)}$ of layer h_l. Each unit $u_i^{(l)}$ of a layer h_l, has an activation function $\sigma_i^{(l)}$ and a bias $b_i^{(l)}$. The output of a unit is the application of the activation function of this unit to its weighted inputs (plus the bias). Output $o_i^{(l)}$ of unit $u_i^{(l)}$ of layer h_l is given by the following equation:

$$o_i^{(l)} = \sigma_i^{(l)}\left(\sum_{j=1}^{n_{l-1}} w_{j,i}^{(l)} \cdot o_j^{(l-1)} + b_i^{(l)}\right), l \geq 1 \qquad \text{Eq. 16}$$

where n_{l-1} is the number of units in layer h_{l-1}.

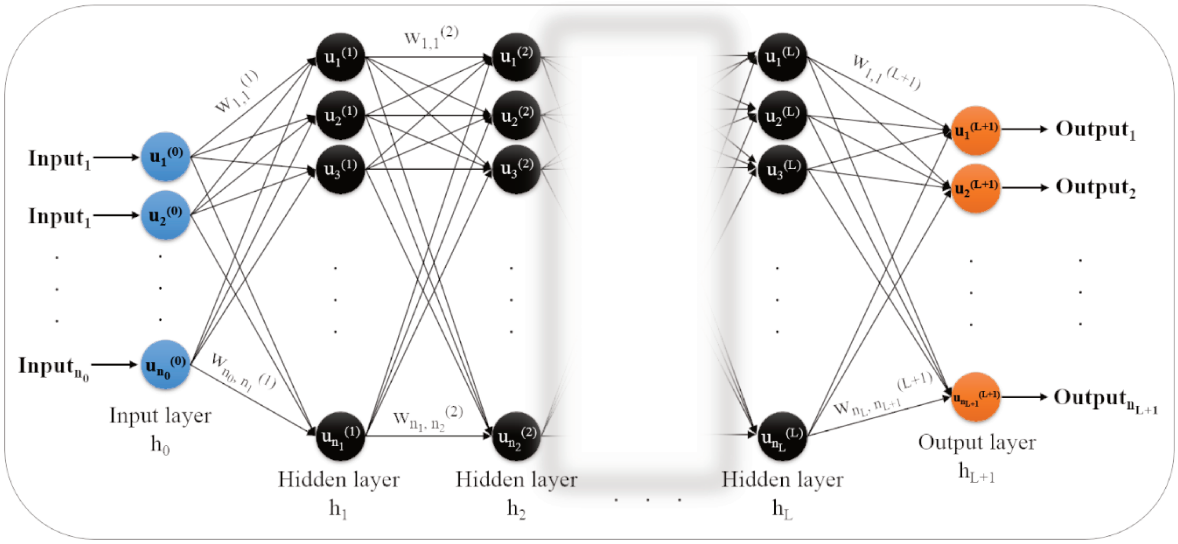


Fig. 15: Feed-forward neural network. Units are organized into layers. All units of a layer are connected to all units of the next layer.

When a feed-forward neural network is used for a regression problem, its input units take the value of the input data. The output units are generally chosen to have a linear activation function. A network is typically "trained" – i.e. its weights, biases, and other parameters are estimated – on a training set. Once trained, it can be used to predict outputs from new inputs. This learning is classically done using the back-propagation algorithm (Rumelhart et al., 1986).

2. Training artificial neural networks

Training an artificial neural network for a regression problem consists in automatically finding the best parameters (weights and biases) of the network so that it correctly captures the relationship between a given set of input and target data, here between articulatory and acoustic features. This requires to have a parallel dataset of corresponding input and target data. An optimization algorithm is then used in order to minimize the difference between the output of the neural network and the target, for a given input from the training data, according to some error criterion.

The most commonly used algorithms for training artificial neural networks are based on the classic back-propagation algorithm (Rumelhart et al., 1986), such as gradient descent or conjugate gradient (Shewchuk, 1994; Rasmussen, 1996), but many other algorithms exist, for instance genetic algorithms (Montana and Davis, 1989; Korning, 1995) or the Levenberg-Marquardt method (Gavin, 2013). However, genetic algorithms rely on random processes and are thus unlikely to be well suited to train deep neural networks which have several hundreds of thousands or millions of parameters to estimate. They might however be used in combination with back-propagation (David et al., 2014). The Levenberg-Marquardt method has proven to be efficient but relies on the computation on the Jacobian matrix of the error function with regard to the network parameters. While this matrix can be computed using similar approach as the back-propagation algorithm, its dimension is equal to $(B * O) \times P$, with B the number of samples in each training batch, O the number of outputs of the network and P the total number of parameters weights and biases. For instance, a network with three hidden layers of 200 units each, which maps 27 articulatory parameters to 25 acoustic parameters would have a Jacobian

matrix of dimension 2500×91025 if the training is performed in batches of 100 samples each. In terms of memory use and computation time, this is too demanding. On the other hand, back-propagation-based algorithms have proven to be efficient and to lead to good solutions. This is why we chose here to use the conjugate gradient (CG) algorithm, which uses back-propagation to estimate the final parameters of the network.

Back-propagation is extensively described in the literature (Bishop and Christopher M. B, 2006). This is an efficient algorithm to compute the gradient of the error function with regards to the network parameters, based on the derivate chain rule. A null gradient meaning that we are at a local minimum of the error function, this gradient gives local information which can be interpreted as the opposite direction (positive or negative) in which we should change each network parameter in order to decrease the error of the network – but this does not give any information about how much we should go in that direction. The opposite of the gradient is thus often called the search direction. The simplest form of optimization based on this gradient is the so-called "gradient descent" (GD), which iteratively updates the parameters $\mathbf{p_i}$ by subtracting from them the gradient of the error function E, multiplied by a factor α, the learning rate:

$$\boldsymbol{p_{i+1}} = \boldsymbol{p_i} - \alpha \cdot \nabla_{\mathbf{p}} E(\boldsymbol{p_i}) \qquad \text{Eq. 17}$$

One main drawback of this simple approach is that since no information is given about how much we should change the parameters, and the learning rate being arbitrarily chosen, we could change them too much and move away from the local minimum, which might even lead to a les optimal local minimum. Alternatively, we could not change them enough so that the training process would be too slow or would be trapped in a local minimum without being able to escape from it. To compensate for this, many variations of the gradient descent method have been proposed such as adapting the learning rate during the training, starting with a high learning rate in order to be able to escape from local minima at the beginning of the training, then reducing it at each iteration in order to stay in the found minima, or by adding a momentum term which will smooth the changes of direction of the gradient from one iteration to another, etc. Most of the time these parameters (learning rate, momentum, etc.) have to be manually set by trials and errors.

An alternative approach to the gradient descent (GD) is the conjugate gradient method (CG), which fundamental basis is detailed in (Shewchuk, 1994). For training an ANN, the conjugate gradient method works by iteratively computing search directions in conjugation with a line search algorithm. The purpose of a line search algorithm is to find the point that minimize a function f along a line, i.e. the search space for a solution minimizing f is restricted to a line. Since there are different versions of the conjugate gradient, the following briefly explains a specific one we implemented for our experiments.

In a first step, the CG algorithm computes the search direction $\mathbf{s_0}$, which is the opposite of the gradient of the error function E with regards to the parameters $\mathbf{p}$, using the initial value of the parameters $\mathbf{p_0}$, as in the GD approach:

$$\boldsymbol{s_0} = -\nabla_{\mathbf{p}} E(\boldsymbol{p_0}) \qquad \text{Eq. 18}$$

It then performs a line search along this direction $\mathbf{s_0}$ to find the point $\mathbf{p_1}$ that minimize the error function and update the network parameters, which can be interpreted as finding the best learning rate α_0 for this iteration when using the GD approach:

$$\boldsymbol{p_1} = \boldsymbol{p_0} + \alpha_0 \mathbf{s_0} \quad \text{Eq. 19}$$

In the line search, the two Wolfe's conditions were used (Wolfe, 1969, 1971). The first one stipulates that α_0 should give sufficient decrease of the error function, with regards to some constant ρ:

$$E(\boldsymbol{p_1}) \leq E(\boldsymbol{p_0}) + \rho(\boldsymbol{p_1} - \boldsymbol{p_0})^T \nabla_{\mathbf{p}} E(\boldsymbol{p_0}) \quad \text{Eq. 20}$$

The second condition stipulates that the choice of α_0 should result in a smaller gradient than the previous one with regards to some constant ε, in order to guarantee that the algorithm moves closer to a local minimum by a non-vanishing amount:

$$\left|\nabla_{\mathbf{p}} E(\boldsymbol{p_0})\right| \leq \varepsilon \left|\nabla_{\mathbf{p}} E(\boldsymbol{p_1})\right| \quad \text{Eq. 21}$$

The line search algorithm uses these two conditions to test if a new point is significantly better than the current point. The line search algorithm used here iterates until an acceptable point that satisfies both conditions is found, and uses quadratic and cubic polynomial fits in order to limit the number of guesses needed (see (Rasmussen, 1996) for more details).

At the next step, the new search direction $\mathbf{s_1}$ is computed using the new gradient values for the parameters $\mathbf{p_1}$ combined with the previous search direction, multiplied by a factor β_1 which can be seen as a variable momentum factor in the GD approach:

$$\boldsymbol{s_1} = -\nabla_{\mathbf{p}} E(\boldsymbol{p_1}) + \beta_1 \mathbf{s_0} \quad \text{Eq. 22}$$

The factors β_i can be computed using different heuristics, which will not be discussed here. In the present study, the Polak-Ribière formula (Polak and Ribiere, 1969) was used (Eq. 23). This formula results from approximating the error function as being locally quadratic, and is recommended among others (Press et al., 1988):

$$\beta_i = max(0, \frac{\nabla_{\mathbf{p}} E(\boldsymbol{p_i})^T \left(\nabla_{\mathbf{p}} E(\boldsymbol{p_i}) - \nabla_{\mathbf{p}} E(\boldsymbol{p_{i-1}})\right)}{\nabla_{\mathbf{p}} E(\boldsymbol{p_{i-1}})^T \nabla_{\mathbf{p}} E(\boldsymbol{p_{i-1}})}) \quad \text{Eq. 23}$$

The line search algorithm is then applied to find the point $\mathbf{p_2}$ that minimizes the error function along the direction $\mathbf{s_1}$. This process is then iteratively repeated, until some convergence criterion is reached, or until the error reaches a maximum number of evaluations, in order to reduce computation time.

3. Difficulties when training deep neural networks

Deep neural networks (DNNs) can be seen as feed-forward neural networks, which have at least two hidden layers. They generally need a specific training algorithm due to this deeper architecture. DNN training using back-propagation is usually a complex task since large initial weights typically lead to poor local minima, while small initial weights lead to small gradients making the training infeasible with many hidden layers (Hinton and Salakhutdinov, 2006). This issue is generally solved by using a pre-training step, in which a specific algorithm is used to

correctly initialize the weights and biases of the network – generally layer by layer – before using a classic back-propagation-based method to fine-tune the final network (Bengio et al., n.d.; Hinton and Osindero, 2006; Hinton and Salakhutdinov, 2006).

For predicting continuous variables, as acoustic parameters, the most common pre-training approach consists in pre-training a generative model called a Deep Belief Network (DBN) which will serve to initialize the neural network to be fine-tuned by back-propagation (Hinton and Osindero, 2006; Hinton, 2010a). This method has been proven to be efficient for the inverse problem of the acoustic-to-articulatory mapping, i.e. predicting the configuration of the vocal tract from an audio signal (Uría, 2011). However, preliminary results showed that using such a pre-training method resulted in poor final solution with our training dataset and we thus proposed another training approach, which will be presented later in this thesis.

VII. Conclusion

In this chapter we briefly introduced speech production and synthesis before focusing on articulatory-based speech synthesis and ways of acquiring articulatory data. Our choice for articulatory-based speech synthesis was motivated by studies exhibiting a somatotopic organization of the speech motor cortex, from which we plan to record neural activity in order to decode speech.

Among all the available methods for recording articulatory data, we made the choice to use electromagnetic articulography (EMA) which presents several advantages with respect to other methods. In particular, EMA allows to record three-dimensional positions from all the main speech articulators – lips, tongue, jaw and soft palate – with high spatial and high temporal resolutions, which are required to capture accurate and fast constrictions of the vocal tract. Moreover, EMA recordings can be done during a sufficiently long period of time (about one or two hours). Even if this time is limited by the fact that glued sensors detach themselves because of salivation, it is still long enough to record large articulatory-acoustic datasets as needed for machine learning techniques to synthesize speech from articulatory data.

While physical approaches that model the vocal tract geometry and acoustic properties can be used to synthesize speech from articulatory data, the approximations needed to solve the mathematical models behind them might limit the sounds they can synthesize. Moreover, once trained, a machine learning model is usually very fast to apply for real-time synthesis, which is a requirement for a brain-computer interface for speech rehabilitation. We thus made the choice to build our articulatory-based speech synthesizer using a machine learning approach. Among the different possible approaches, we chose to perform a continuous mapping of the articulatory data into acoustic trajectories without passing through an intermediate phonetic representation as in articulatory recognition. This choice was motivated by the fact that continuous mapping ensure a minimal delay between the articulatory input and the output sounds, as opposed to articulatory recognition approaches for which the synthesis cannot occur before a phonetic unit, such as a word, is fully recognized.

In this thesis, we thus chose to perform articulatory-based speech synthesis using a machine-learning based approach. In particular, we chose to compare two methods, Gaussian mixture models and deep neural networks, which we present in details in part 3.

Part 2: Goal of the thesis

As shown in the previous chapters, to date, there has not yet been any demonstration of an open-vocabulary BCI able to reconstruct continuous intelligible speech in real-time (Guenther et al., 2009). The goal of this thesis was thus to develop several aspects toward such proof of concept, with patients undergoing awake surgery for a tumor removal, that had preserved brain areas and were able to speak. In particular, this thesis had three aims:

1) To develop a speech synthesizer that produces fully intelligible speech from articulatory data in real-time, that has few control parameters and that is as robust as possible to noisy control parameters.
2) To investigate the feasibility of decoding speech and articulatory features from neural activity essentially recorded in the speech motor cortex.
3) To consider ethical issues that arise with the development and use of brain-computer interfaces.

The first aim was thus to develop a speech synthesizer that could produce intelligible speech from articulatory movements, i.e. that converts movements of the main speech articulators, such as the lips or the tongue, into an audible speech signal. The choice of synthesizing speech from articulatory movements, rather than another representation – for example a phonetic sequence such as in text-to-speech synthesis – was motivated by several studies showing that the speech sensorimotor cortex exhibits a topographic organization mapping the different articulators involved in speech production (Penfield and Boldrey, 1937; Guenther, 2006; Pulvermüller et al., 2006; Grabski et al., 2012; Tate et al., 2014), and furthermore motivated by a more recent study showing that during speech production, the activity of the speech sensorimotor cortex is tuned to the articulatory properties of the produced sounds but not to their acoustic properties (Cheung et al., 2016). While synthesizing speech from articulatory data had been already investigated in several studies (Maeda, 1982; Sondhi and Schroeter, 1987; Kello and Plaut, 2004; Birkholz et al., 2006; Toda et al., 2008), there was no demonstration of intelligible synthesis from EMA data for the French language, and it remained unknown whether a given articulatory-based speech synthesizer built from articulatory-acoustic data obtained in one particular reference speaker could be controlled in real time by any other speaker to produce intelligible speech. Thus, our aim was not only to develop a speech synthesizer that could synthesize intelligible speech from articulatory data, but also to make this synthesizer run in real time and controlled by other subjects, in different experimental conditions, as a first step toward brain control in a BCI paradigm. **Part 3** presents results for this first aim, with **chapter 3** focusing on the acquisition of an articulatory-acoustic corpus, **chapter 4** focusing on speech synthesis from articulatory data in offline condition, and **chapter 5** focusing on the control of this synthesizer by different subjects in closed-loop conditions.

The second aim was to investigate the feasibility of decoding speech and articulatory features from neural activity. The second part of this thesis was thus dedicated to the development of methodological tools – in majority software tools – since brain-computer interfaces were a newly developed research axis in the team. In particular, the first step was to develop a per-operative method to localize the speech-related brain areas using

electrocorticographic (ECoG) recordings. Indeed, it was envisioned to use micro-electrode arrays (MEAs) to record neural activity since very promising results in BCI have been achieved using MEAs (Collinger et al., 2013; Ifft et al., 2013; Wodlinger et al., 2014). However, such MEAs can only cover a minor part of the brain cortex so that their positioning has to be optimized in order to record neural activity that is as informative as possible for decoding speech and articulatory features. A goal of this thesis was thus to develop a tool to precisely localize speech-related brain areas before the implementation of a MEA, directly during the surgery. In a second step we performed initial tests to decode neural data recorded from the speech-related brain areas to predict articulatory and speech features from neural activity. We investigated in particular the case of speech intention detection, i.e. to predict if the patient is speaking or intending to speak from the neural activity. While speech BCIs would mostly benefit aphasic patients with preserved cortical areas, such as in the locked-in syndrome, these developments were made using neural activity recorded from patients undergoing awake surgery for a tumor removal that had no speech disorder. These results are presented in **Part 4**. **Chapter 6** presents the per-operative localization of speech-related brain areas while **chapter 7** focuses on preliminary decoding of neural activity.

Finally, my thesis was also an opportunity to conduct ethical reflections on the development and use of brain-computer interfaces. These considerations are presented in **Part 5**.

Part 3: Thesis result 1 – Articulatory-based speech synthesis for BCI applications

As previously motivated in Chapter 1, the goal of this thesis was to set the ground for a Brain Computer Interface (BCI) for speech restauration, in which neural activity is recorded from the speech motor cortex and decoded in control parameters for an articulatory-based synthesizer. This choice was highly motivated by recent studies showing that the speech motor cortex activity exhibits features correlated to articulatory features of speech during speech production (Bouchard et al., 2013; Cheung et al., 2016). Moreover, while speech synthesis from articulatory data can be achieved in several ways and using different types of articulatory data, we motivated in **Chapter 2** our choice to use a machine learning approach in which a large articulatory-acoustic dataset recorded using electromagnetic-articulography (EMA) is used to train a statistical model mapping articulatory trajectories into acoustic parameters.

In **Chapter 3**, we thus present two articulatory-acoustic datasets that were used to synthesize speech from articulatory data. The first one, PB2007, was an existing corpus, while the second one, BY2014, was a new corpus, specifically recorded for this study.

In **Chapter 4**, we present how this corpus was used to build an articulatory-based speech synthesizer which converts articulatory parameters into an audible speech signal. Two approaches were compared: a state-of-the-art approached based on Gaussian mixture models, and our approach, which used deep neural networks.

In **Chapter 5**, we present how the same articulatory-based speech synthesizer can be controlled by several subjects in real-time, from articulatory data recorded in silent speech condition. Such condition was chosen to be as close as possible to a BCI paradigm for speech rehabilitation.

Chapter 3: The PB2007 and BY2014 articulatory-acoustic corpora

I. Introduction

In order to convert articulatory trajectories into acoustic parameters using machine learning techniques, a first step consists in recording a large articulatory-acoustic database, in which articulatory data from a speaker is recorded synchronously with the produced audio speech signal. The articulatory-acoustic dataset could be chosen as large as possible since it only has to be recorded once in a single subject. However, EMA requires the whole dataset to be recorded at once since EMA sensors cannot be placed at the exact same positions between two different sessions. This limits the recording time to a maximum of about two hours since sensors can detach because of salivation, and in order to ensure sufficient comfort of the recorded subject. There are few EMA dataset available publicly, notably the MOCHA-TIMIT dataset (http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html), containing about 460 sentences recorded from 4 different subjects synchronously with EMA data, and the mngu0 dataset (http://www.mngu0.org/), consisting of more than 1,300 phonetically diverse utterances recorded from a single British English speaker. However, these corpora are generally in English, while here we wanted to develop a speech synthesizer for French. Only one French corpus was available in our lab, the PB2007 corpus (Ben Youssef et al., 2009). We first used this corpus in order to test different strategies to perform synthesis from articulatory data. Then we extended that dataset by recording another corpus, the BY2014 corpus.

In this chapter, we will thus first describe the PB2007 articulatory-acoustic corpus, and then the BY2014 corpus.

II. The PB2007 corpus

1. Articulatory data acquisition and parametrization for the PB2007 corpus

The PB2007 was an already existing corpus (Ben Youssef et al., 2009) recorded from a native French speaker using the Carstens 2D EMA system (AG200). Six coils were glued on the tongue tip, blade, and dorsum, as well as on the upper lip, the lower lip and the jaw (Fig. 16-A). Extra coils attached to the upper incisors and to the nose served as references to compensate for head movements in the midsagittal plane. Articulatory data was low-pass filtered at 20 Hz and down-sampled from 200 Hz to 100 Hz. **Fig. 16**-B shows an example of recorded articulatory trajectories projected in the midsagittal plane, with the corresponding audio signal.

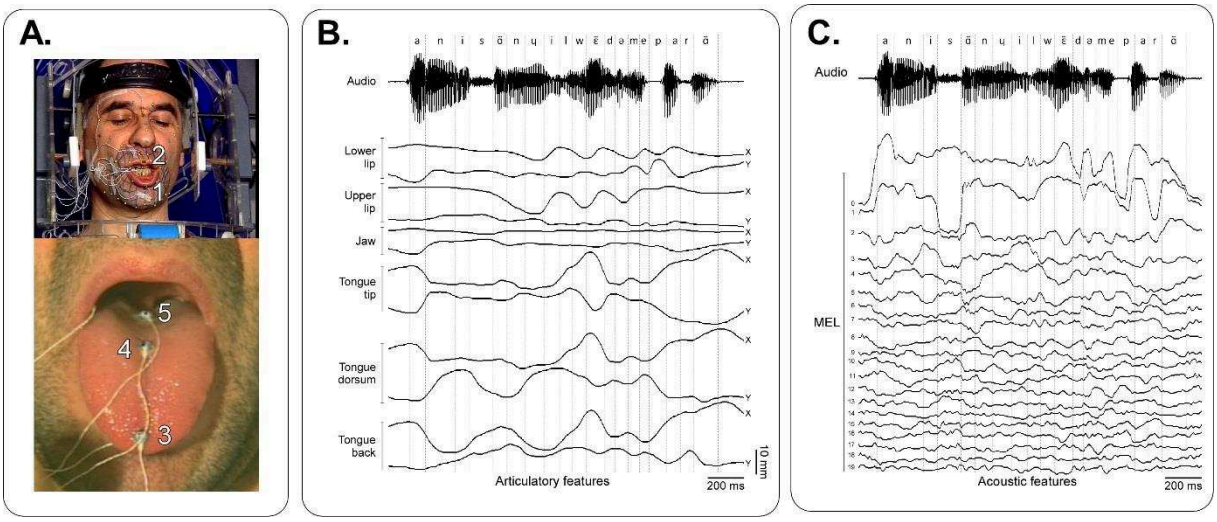


Fig. 16: PB2007 articulatory and acoustic data. A – Positioning of the sensors on the upper lip (1), lower lip (2), tongue tip (3), tongue dorsum (4), and tongue back (5). The jaw sensor was glued at the base of the incisive (not visible in this image). B – *Articulatory signals and corresponding audio signal for the sentence "Annie s'ennuie loin de mes parents" ("Annie gets bored away from my parents"). For each sensor, the horizontal caudo*-rostral X and below the vertical ventro-dorsal Y coordinates projected in the midsagittal plane are plotted. Dashed lines show the phone segmentation obtained by forced-alignment. C – Acoustic features (20 mel-cepstrum coefficients - MEL) and corresponding segmented audio signal for the same sentence as in B.

2. Acoustic data acquisition and parametrization for the PB2007 corpus

The acoustic speech signal was recorded at 22,050 Hz synchronously with the articulatory data, and down-sampled at 16 kHz.

Its spectral content was parameterized by 20 complex mel-cepstrum (MEL) coefficients computed every 10 ms (hence a 100 Hz sampling matching the articulatory data acquisition frequency) from a 25-ms sliding window (**Fig. 16**-C). The computation of these MEL coefficients was done using the Speech Processing ToolKit (SPTK) mcep tools (Tokuda et al., 2014). These 20 coefficients efficiently represent the spectral envelope of speech and can be converted back into audible sounds by building a so-called Mel Log Spectrum Approximation (MLSA) filter (Imai, 1983). This approach is based on the source-filter model of speech production, which models the speech signal as a convolution of a sound source (e.g., the glottal activity) with a linear acoustic filter representing the vocal tract. In the present MLSA model, a set of M mel-cepstrum coefficients $c_\alpha(m)$ represent the vocal tract filter H(z) for each audio signal window, as follows:

$$H(z) = \exp \sum_{m=0}^{M} c_\alpha(m).\tilde{z}^{-m},$$

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

The coefficient α is chosen so that the mel-scale becomes a good approximation of the human sensitivity to the loudness of speech (here α = 0.41). The mel-cepstral coefficients $c_\alpha(m)$ are approximated using the Newton-Raphson method for numerically solving equations, and linearly combined to obtain the MLSA filter coefficients. This filter is then excited with a source signal representing the glottal activity (i.e. vibration of the vocal folds) in order to reconstruct the corresponding speech. Such excitation signal is generally designed by extracting

the pitch from the original audio signal, and then generating white noise for non-voiced segments, and a train of pulses for voiced segments, which period matches the pitch.

3. Content of the PB2007 corpus

Transcription of the audio signals was first done manually in naturally written text, then translated into phone sequences using LLiaPhon phonetizer (Bechet, 2001), and finally manually corrected. Phone sequences were then automatically aligned on audio files using a forced-alignment procedure, based on a standard speech recognition system.

The PB2007 database contained about 15 minutes of speech after removing the periods of silence. This database was composed of 1108 items, including all isolated vowels, all vowel-consonant-vowel sequences (VCVs) with identical start and end vowel, many isolated words, and some phonetically balanced sentences. This resulted in 5,115 phones in total. The distribution of the 34 different phonetic classes used to describe French language in this corpus is shown in Fig. 17-A. Phone frequency ranged from 348 occurrences for the phone /a/ to 38 occurrences for /h/. The distribution of the articulatory data points in the midsagittal plane is represented in Fig. 17-B. The jaw was the articulator with the smallest movement amplitude (about 10mm), followed by the upper lip (about 15mm), the lower lip (about 20mm), and finally the tongue had the highest amplitude of movement (about 30mm for each sensor).

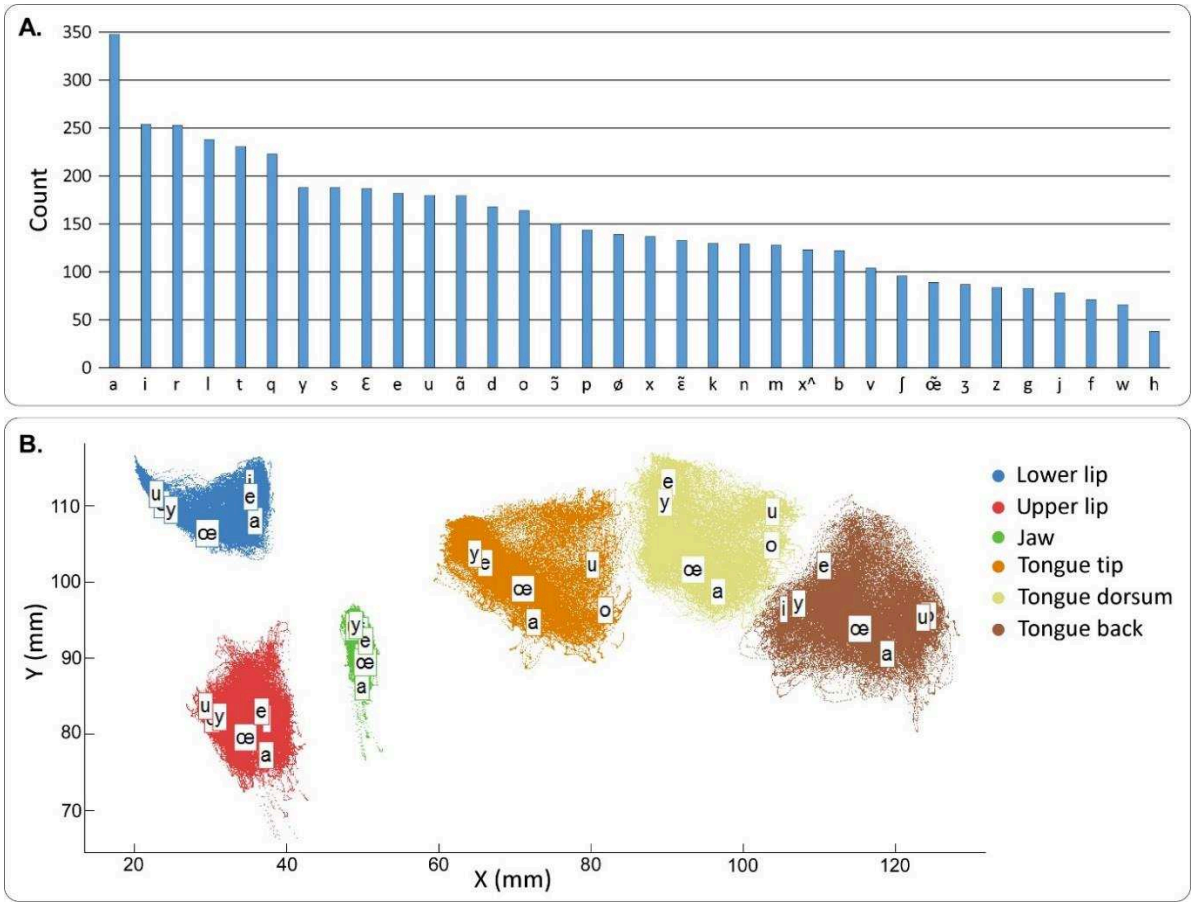


Fig. 17: PB2007 articulatory-acoustic database description. A – Occurrence histogram of all phones of the articulatory-acoustic database. Each bar shows the number of occurrence of a specific phone in the whole corpus. B – Spatial distribution of all articulatory data points of the database (silences excluded) in the midsagittal plane. The positions of the different sensors are plotted with different colors. The labeled positions correspond to the mean position for the 7 main French vowels.

III. The BY2014 corpus

In a preliminary step, the PB2007 corpus was first used to build an articulatory-based speech synthesizer, which lead to promising results (see "Conclusion on the PB2007 synthesis"). However, preliminary results suggested that better synthesis accuracy could be achieved by recording a larger articulatory-acoustic database including more full sentences instead of isolated words, as well as adding at least one sensor on the soft palate in order to discriminate nasal and non-nasal phones. The BY2014 corpus was thus recorded for this purpose.

1. Articulatory data acquisition and parametrization for the BY2014 corpus

The articulatory data was recorded using the electromagnetic articulography (EMA) NDI Wave system (NDI, Ontario, Canada), which allows three-dimensional tracking of the position of small coils with a precision of less than a millimeter (while the EMA system used for the PB2007 corpus only allowed to record 2D positions of the coils in the midsagittal plane). Nine such 3D coils were glued on the tongue tip, dorsum, and back, as well as on the upper lip, the lower lip, the left and right lip corners, the jaw and the soft palate (**Fig. 18**-A). This configuration was chosen for being similar to the ones used in the main publicly available databases, such as MOCHA (http://www.cstr.ed.ac.uk/ research/projects/artic/mocha.html) and mngu0 (Richmond et al., n.d.), and in other studies in articulatory-based synthesis (Toda et al., 2008) or articulatory-to-acoustic inversion (Uria et al., 2011). It is similar to that of the PB2007 dataset and allows to capture well the movements of the main articulators while avoiding to perturb the speaker too much: 3 coils on the tongue give information on back, dorsum and apex while 4 coils on lips give information on protrusion and rounding, and we considered that one sensor was enough for the jaw since it is a rigid articulator, and one for the soft palate since it has mostly one degree of freedom. An additional 6D reference coil (which position and orientation can be measured) was used to account for head movements and was glued behind the right ear of the subject. To avoid coil detachment due to salivation, two precautions were taken to glue the sensors. First, the tongue and soft palate sensors were glued onto small pieces of silk in order to increase contact surface, and second, the tongue, soft palate and jaw surfaces were carefully dried using cottons soaked with 55% green Chartreuse liquor. The recorded sequences of articulatory coordinates were down-sampled from 400 Hz to 100 Hz. **Fig. 18**-B shows an example of recorded articulatory trajectories projected in the midsagittal plane, with the corresponding audio signal. The vibration of the vocal folds was as well recorded using an electroglottograph (EGG), but this signal was not used in the present study.

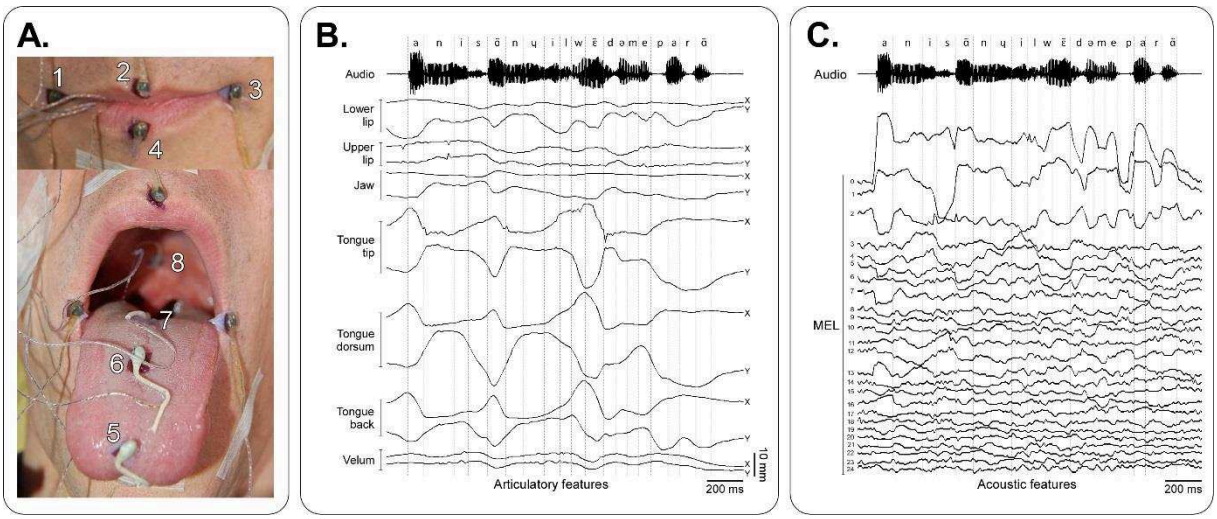


Fig. 18: BY2014 articulatory and acoustic data. A – Positioning of the sensors on the lip corners (1 & 3), upper lip (2), lower lip (4), tongue tip (5), tongue dorsum (6), tongue back (7) and velum (8). The jaw sensor was glued at the base of the incisive (not visible in this image). B – *Articulatory signals and corresponding audio signal for the sentence "Annie s'ennuie loin de mes parents" ("Annie gets bored away from my parents"). For each sensor, the horizontal caudo*-rostral X and below the vertical ventro-dorsal Y coordinates projected in the midsagittal plane are plotted. Dashed lines show the phone segmentation obtained by forced-alignment. C – Acoustic features (25 mel-cepstrum coefficients - MEL) and corresponding segmented audio signal for the same sentence as in B.

2. Acoustic data acquisition and parametrization for the BY2014 corpus

The acoustic speech signal was recorded at 22,050 Hz synchronously with the articulatory data. Its spectral content was parameterized by 25 MEL coefficients computed every 10 ms (hence a 100 Hz sampling matching the articulatory data acquisition frequency) from a 23-ms (512 samples) sliding window using the SPTK mcep tools (**Fig. 18**-C). The choice was made to increase the number of MEL coefficients with regards to the PB2007 corpus in order to have a more precise parametrization of the spectral content which might lead to a better synthesis quality. The same reason motivated the choice to keep the original audio sampling rate of 22,050 Hz instead of decreasing it to 16 kHz. Note that the α coefficient used for the MLSA depends on the audio sampling frequency and was thus equal here to 0.455 instead of 0.41.

3. Content of the BY2014 corpus

As for the PB2007 corpus, transcription of the audio signals was first done manually in naturally written text, then translated into phone sequences using LLiaPhon phonetizer, and finally manually corrected ; and phone sequences were then automatically aligned on audio files using a forced-alignment procedure, based on a standard speech recognition system.

The final database contained more than 45 minutes of speech after removing the periods of silence, which was about three times longer than the PB2007 corpus. This database was composed of 925 items of variable length, including all isolated vowels, all vowel-consonant-vowel sequences (VCVs) with identical start and end vowel, phonetically balanced sentences, and many other sentences extracted from articles of the French newspaper "Le Monde". This resulted in 18,828 phones in total, which was almost four times more than the PB2007 corpus. The distribution of the 34 different phonetic classes used to describe French language in this

corpus is shown in Fig. 19-A. Phone frequency ranged from 1,420 occurrences for the phone /a/ to 27 occurrences for /ɲ/. The distribution of the articulatory data points in the midsagittal plane is represented in **Fig. 19**-B. The velum was the articulator with the smallest movement amplitude (less than 10mm), followed by the jaw and the upper lip (about 10mm), the lower lip (about 20mm), and finally the tongue had the highest amplitude of movement (about 30mm for each sensor).

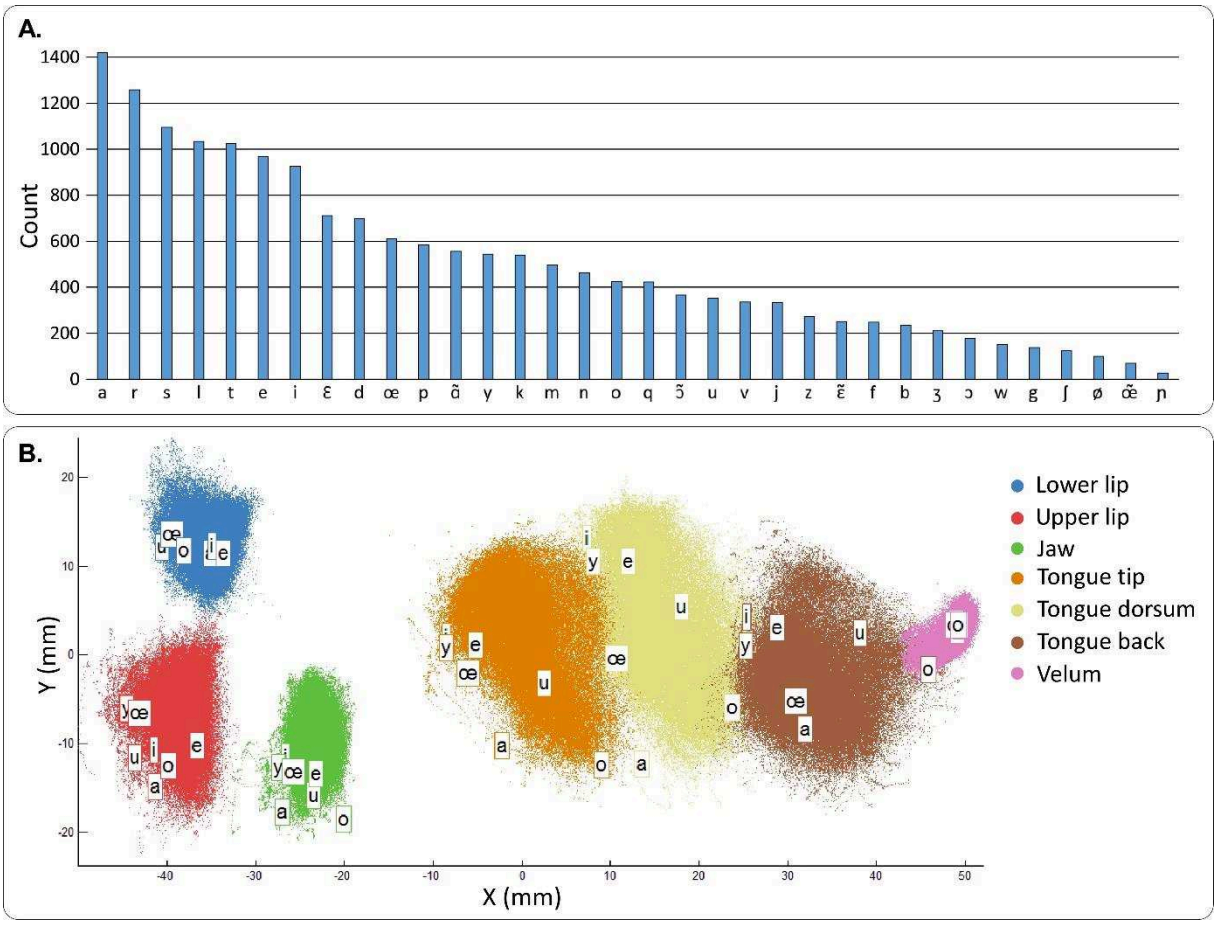


Fig. 19: BY2014 articulatory-acoustic database description. A – Occurrence histogram of all phones of the articulatory-acoustic database. Each bar shows the number of occurrence of a specific phone in the whole corpus. B – Spatial distribution of all articulatory data points of the database (silences excluded) in the midsagittal plane. The positions of the different sensors (except corner lips) are plotted with different colors. The labeled positions correspond to the mean position for the 7 main French vowels.

The whole BY2014 corpus was made publicly available and can be downloaded at https://zenodo.org/record/154083 (Bocquelet et al., 2016b).

Chapter 4: Articulatory-based speech synthesis

I. Introduction

In the previous chapter we presented two articulatory-acoustic datasets, the PB2007 corpus, which was an already existing corpus, and the BY2014 corpus, which was a new corpus that we recorded to extend the PB2007 corpus. Such synchronous articulatory-acoustic data can then be used to automatically compute (or "train") mathematical models in order to convert (or "map") new articulatory trajectories into the corresponding acoustic parameters, the so-called articulatory-to-acoustic mapping. The types of mathematical models as well as the methods to train them are numerous and are part of an entire research field: machine learning. At the beginning of this thesis, state-of-the-art methods for articulatory-based speech synthesis (based on machine learning) were using Gaussian Mixture Models (GMM), while Deep Neural Networks (DNN) were showing many promising results in other fields. This motivated our choice to test and compare both approaches.

In this chapter, we thus describe how articulatory trajectories were mapped to acoustic parameters using a Deep Neural Network (DNN), and compare our results with a state-of-the-art approach based on Gaussian Mixture Models (GMMs).

II. Articulatory-to-acoustic mapping

The articulatory-to-acoustic mapping was performed both using GMMs and DNNs. Fundamental basis of GMMs and DNNs models and their training are detailed in **Chapter 2**. Here we describe the choices of GMMs and DNNs parameters that we made to perform the articulatory-to-acoustic mapping.

1. GMM-based mapping

Gaussian Mixture Models (GMMs) have been previously used to predict acoustic parameters from EMA data in particular using the trajectory GMM approach proposed by Toda et al. that we presented in Chapter 2, which takes into account the data dynamics (Toda et al., 2008). Here we considered this approach as a gold-standard to which we compared our approach using deep neural networks (DNNs).

a. Choice of GMM hyper-parameters

One advantage of using the trajectory GMM for regression is that it has very few hyper-parameters (i.e. parameters that are not learned during the training phase but are chosen a priori).

Number of components. The first parameter is the number M of components in the mixture. While different approaches have been proposed to try to automatically estimate the

best number of components, such as modification of the EM algorithm (Huang et al., 2013), we chose here to train different mixtures with different numbers of components (from 16 to 256) in order to observe how this affect the synthesis quality, and observe any overfitting effect (when the model has many trained parameters, it overfits the training data, thus not being able to correctly predict unknown data).

Order of the derivatives. In order to take into account the dynamic properties of acoustic parameters, the trajectory GMM includes the derivatives of the acoustic features. Thus the order of the derivatives is another hyper-parameter. Here we chose to only add the first derivatives to the acoustic features vector, as defined in (Eq. 11). Indeed, while this is not reported in the present manuscript, preliminary results showed no significant improvement when using higher order derivatives, such as the second order derivatives defined in (Eq. 12).

Training parameters. Finally, additional hyper-parameters are relative to the training phase. First, the EM algorithm needs an initial estimation of the searched parameters, that is to say an initial estimate of the mean vector and covariance matrix of each Gaussian component. For that purpose, we used the k-means algorithm in order to cluster the training data in as many groups as desired Gaussian components. The centroid and covariance of each cluster was then chosen as a first estimation of each Gaussian component. K-means is an iterative algorithm that has its own set of hyper-parameters. Here we used a maximum of 250 iterations, and two replicates (the k-means algorithm needs initial centroids for each cluster. In the first replicate, they are chosen randomly in the data. Then in the second replicate, the result from the first one is used to initialize the centroids). The EM algorithm also provides the possibility to use full or diagonal covariance matrices. Here we used diagonal covariance matrices, as suggested by (Toda et al., 2008).

The training of the GMM was done in a 5-fold cross-validation process, so that the original data was randomly divided into 5 chunks, 4 of which were used for training, while the remaining one was used for testing, which was repeated 5 times in order to test all partitions.

b. Implementation details of the trajectory GMM

All the code for trajectory GMM was implemented in MATLAB. The EM algorithm for fitting a GMM to data and the original code for the k-means algorithm were from the Statistics Toolbox. We used the patch from Da Kuang, which accelerates the computation of k-means without changing the final result (http://math.ucla.edu/~dakuang/software/kmeans3.html).

2. DNN-based mapping

In chapter 2 we introduced the fundamental basis of deep neural networks (DNNs) as well as the conjugate gradient method to fit their parameters. We presented as well the issues that arise when training such deep architecture, resulting in a poor gradient propagation. We mentioned as well the existence of pre-training methods to avoid this issue, and in particular the Deep Belief Network approach, which was used for the inverse acoutic-to-articulatory problem of predicting the articulatory configuration from the audio signal (Uria et al., 2011). However, preliminary results showed that using such a pre-training method resulted into poor final solution with our training dataset and we thus proposed another training approach.

a. Proposed approach for training DNNs for regression

We trained our network using the previously described conjugate gradient algorithm (see Chapter 2). However, instead of directly training the whole network – which did not converge – we chose to add the different layers successively (**Fig. 20**). During the first step, the network was only composed by the input layer h_0, the first hidden layer h_1, and the output layer h_{L+1}, temporarily connected to the h_1 layer (**Fig. 20**-A). This initial network was randomly initialized then fine-tuned using the conjugate gradient algorithm, before deleting the temporary output layer (**Fig. 20**-B). Then the next layer was added so that the new network was now composed by the input layer h_0, the first two hidden layers h_1 and h_2, and the output layer h_{L+1}, temporarily connected to the h_2 layer (**Fig. 20**-C). The weights from the input layer h_0 to the first hidden layer h_1 were those obtained at the previous step and the other weights were randomly initialized. The conjugate gradient algorithm was then applied to this network for fine-tuning, before removing the temporary output layer (**Fig. 20**-D). This process was repeated until all the hidden layers were added (**Fig. 20**-E). While this process is not as computationally efficient as doing a pre-training layer-by-layer such as when training a Deep Belief Network, this allowed good and fast convergence of the training process.

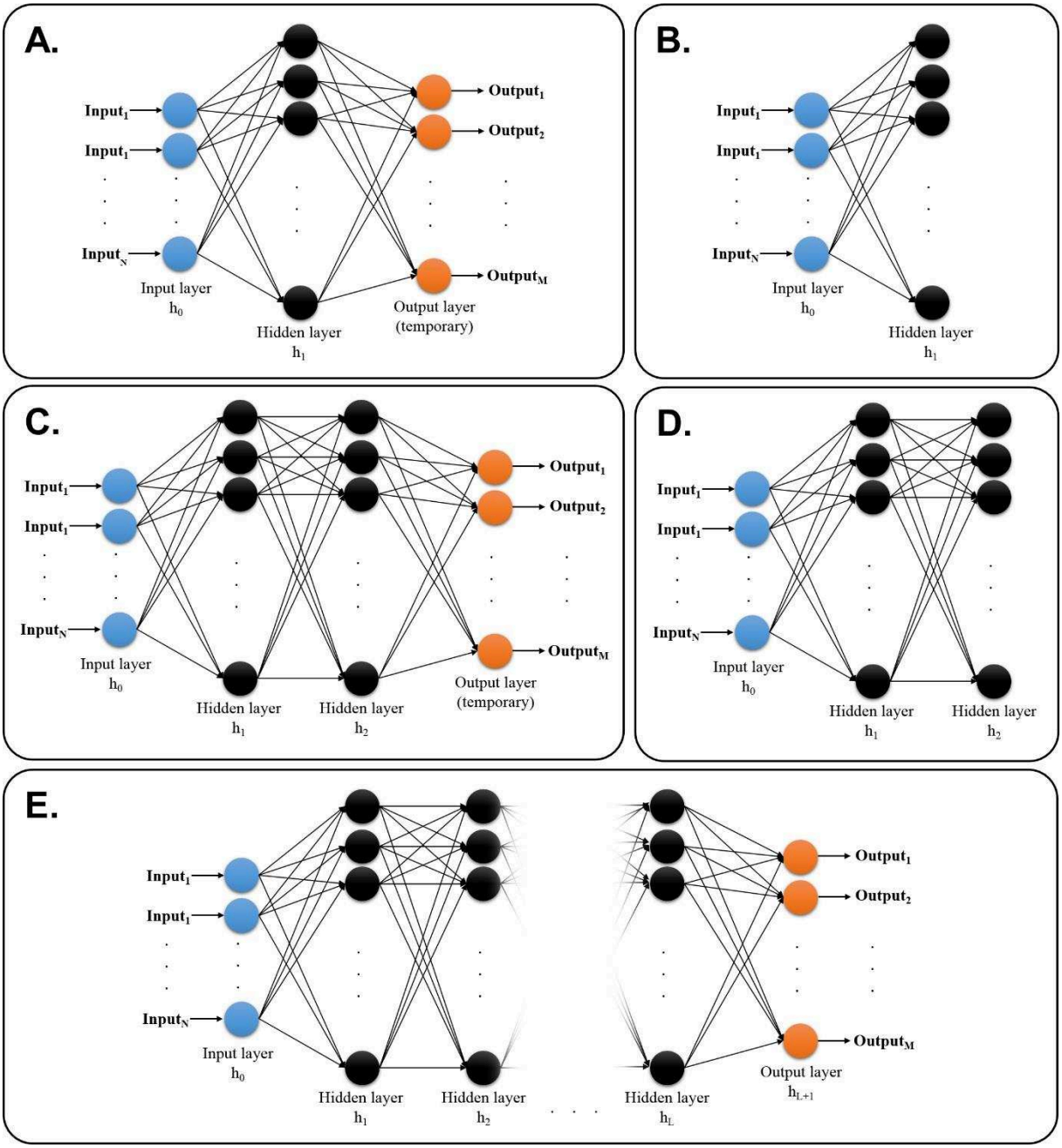


Fig. 20: Proposed approach for training DNNs for regression. A - During the first step, the network was only composed by the input layer h_0, the first hidden layer h_1, and a temporary output layer, connected to the h_1 layer. This initial network was randomly initialized then fine-tuned using the conjugate gradient algorithm. B – The temporary output layer was then deleted. C - The next layer was then added so that the new network was now composed by the input layer h_0, the first two hidden layers h_1 and h_2, and a new temporary output layer, connected to the h_2 layer. The weights from the input layer h_0 to the fist hidden layer h_1 were those obtained at the previous step and the other weights were randomly initialized. The conjugate gradient algorithm was then applied to this network for fine-tuning. D – The temporary output layer was then deleted. E - This process was repeated until all the hidden layers were added.

b. Choice of DNN hyper-parameters

Network architecture. The whole DNN architecture is defined by a set of hyper-parameters: the number of layers, the number of units in each layer (which can be different from one layer to another one), and the activation functions of each unit (which can also be different for each unit, and have its own set of hyper-parameters). Methods have been proposed

to automatically optimize the number of layers and units, such as starting with an initially big networks and dropping units during the training process (Lecun et al., 1990; Zeiler and Fergus, 2012; Hinton, 2014), or using a constructive approach which successively adds new units or layers (Lengellé and Denoeux, 1996). However, as for trajectory GMMs, we chose to train several DNNs with different and manually-set architectures in order to observe the impact on the synthesis quality, and overfitting effects. Here we chose to only consider a DNN with the same number of units in all hidden layers, and all hidden units having the same activation function.

Activation function. Many different activation functions have been proposed in the literature, such as the logistic sigmoid function, the hyperbolic function, the step function, the rectified linear function, the leaky rectified linear function, etc. In fact, any differentiable function could be used. Sigmoid-like function were originally chosen for their step-like shape which could be of advantage when discriminating values for a classification problem. In this study, two different activation functions for the hidden layer were chosen: the logistic sigmoid function (**Fig. 21**-A), and the rectified linear function (**Fig. 21**-B). The logistic sigmoid function is a widely used activation function, and was thus used on the PB2007 corpus for preliminary experiments. However the rectified linear function was suggested by several studies as being more efficient in terms of convergence and classification results (Hinton, 2010b), and was thus used on the new BY2014 corpus. In practice, preliminary experiments not reported here showed indeed a significant improvement on the convergence of the algorithm which needed less iterations and had a steeper learning curve, and less computation time per iteration (the rectified linear function being a piecewise linear function, while the logistic sigmoid includes an exponential term). However, no significant difference was observed in the final error given by the objective function (this was as well observed when using the hyperbolic or leaky rectified linear functions). In order to shorten training times, the rectified linear function was thus preferred to the sigmoid function in future experiments, and thus on the BY2014 corpus.

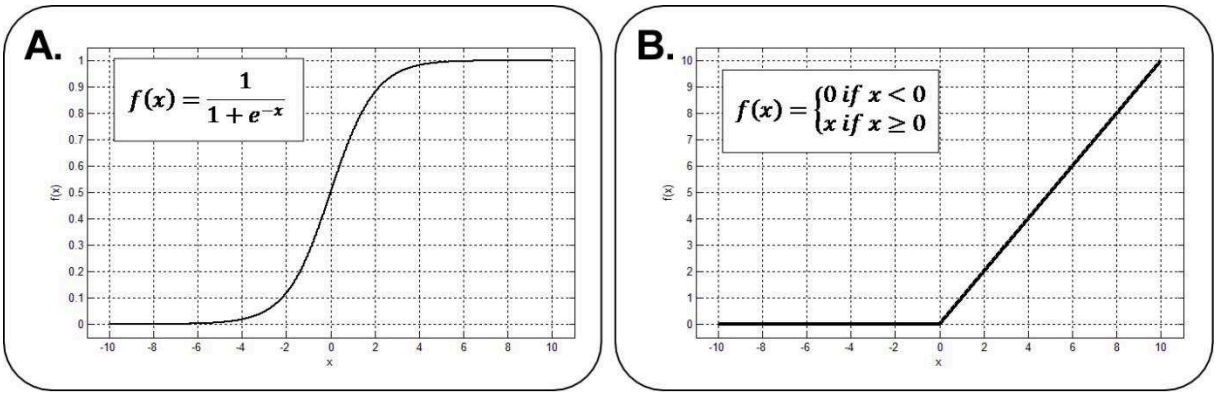


Fig. 21: Activation functions of the neural network. A – Logistic sigmoid function. B –Rectified linear function.

Weights initialization. The training of a DNN being an iterative process, additional hyper-parameters are chosen in order to correctly initialize the weights and biases. This initialization step is of importance since a bad initialization can lead to a poor optimization result. As suggested in (Glorot and Bengio, 2010), bias were null initialized, and weights $w_{i,j}^{(l)}$ from layer h_{l-1} to layer h_l were randomly initialized from a uniform distribution U which interval depended on the activation function σ and the number of units in both layers:

$$W^{(l)} \sim U\left[-\sqrt{\frac{6}{n_{l-1}+n_l}}, +\sqrt{\frac{6}{n_{l-1}+n_l}}\right] \quad \text{Eq. 24}$$

While other different initialization methods have been proposed (Nguyen and Widrow, 1990; Osowski, 1993; Yam and Chow, 2000), in practice choosing the weights in that way leads to good convergence and thus there was no need for further optimization.

Error criterion. Another hyper-parameter is the choice of the error criterion, i.e. the function that measures the difference between the current output of the network, and the target output from the training data, which has to be minimized. The most commonly used error criterion when predicting continuous values is probably the mean squared error (MSE):

$$MSE(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x_i})^2 \quad \text{Eq. 25}$$

With $\boldsymbol{x}$ the target output vector and $\hat{\boldsymbol{x}}$ the current output vector, and n their number of features.

Many other error criterions can be defined, such as using the absolute differences instead of the squared ones, etc. In practice, no clear difference was observed in the final synthesis, so that the MSE was chosen as the error criterion for all trainings.

Batch training. Another hyper-parameter is the choice of how will the networks parameters be updated. For instance, when using gradient descent, the update could be done on a sample-by-sample basis, i.e. the gradient could be computed for one sample to order the parameters, then computed on the next sample to update the parameters, etc. Another way could be to compute the gradient for all samples, then use the mean of all these gradient as the gradient that will be used to update the network parameters. An intermediary and efficient solution is to cut the training data into small "batches" (Hinton, 2010a), and update the network parameters once for each batch using the mean gradient on this batch, which is the solution we used in this work. The data was cut into 100 batches, which consisted of randomly picked samples. When the optimization process iterated over all the batches (i.e. did one "epoch"), new batches were randomly picked, so that from one epoch to another, the batches were never the same.

Training parameters. All the remaining hyper-parameters are those of the training algorithm used. In our case, we used the Polack-Ribière conjugate gradient method as previously described. In practice, the conjugate gradient, when compared with gradient descent, showed the advantage that parameters were not highly dependent on the training dataset, but rather that we could design a set of parameters that would work in most cases. Here, we limited the number of line-searches – i.e. the total number of iterations per update – to 3, and to an overall of 20 function evaluations, which means that the conjugate gradient and the line search algorithms would stop iterating if the error function was computed more than 20 times, ensuring a good compromise between convergence and computation time, and avoiding infinite loops because of numerical inaccuracies. For the constants of the Wolfe's conditions, we chose ρ=0.1 and ε=0.5.

The purpose of the other hyper-parameters of the training are to avoid overfitting of the neural network. Indeed, a neural network with enough units and layers could learn any arbitrary

function of the inputs (Bishop and Christopher M. B, 2006). However, there is a high chance that such a network would be quite inefficient at generalizing, i.e. predict the outputs corresponding to inputs that were not part of the original training dataset. This issue is called overfitting. To avoid overfitting, a widely used approach is called early stopping. In that case, the original training data is divided into two set: a training set that will actually be used to compute the gradient using back-propagation and update the network parameters, and a validation set that will act as new unseen data, which will not be used to update the network parameters, but only to compute the error of the network on this set. During the actual learning stage of the network parameters, the error should decrease on both the training and the validation set. When the network is overfitting, the error still decreases on the training set – since this is the purpose of the optimization algorithm, but generally increases or stabilizes on the validation set. Early stopping consists in stopping the training of the network if there was no improvement on the validation set during several iterations. Here, since the optimization is performed in a batch fashion, i.e. the training data is cut into batches, and the optimization is performed batch by batch instead of feeding all the training data at a time, the error is not strictly decreasing, but slightly jittering. Thus, we chose to stop training if 20 consecutive iterations did not show any improvement on the validation set. Note that this batch training also reduces overfitting by presenting different data to the network at each optimization step, avoiding to minimize the error on a fixed dataset. In practice, 80% of the available data were used for the training, 10% were used for validation, and the remaining 10% for testing. This was done using a 5-fold cross-validation process, so that the original data was randomly divided into 5 chunks, 4 of which used for training, the remaining chunk being used for validation and test, which was repeated 5 times in order to test all partitions – each time with a new network with identical architecture.

c. Implementation details of the DNNs

All the artificial neural networks procedures were done using custom-made optimization tools written in C++. High care was given in order to optimize computation speed: smart memory management was used in order to avoid numerous data transfers between processes, intensive multi-threading was implemented to parallelize and accelerate computations, and formulas were expressed in matrix form when possible, in order to take advantage of fast matrix computation libraries like Eigen (http://eigen.tuxfamily.org). The conjugate gradient code was adapted from the Matlab implementation of the DRToolBox (Rasmussen, 1996). The final framework is highly flexible so that new types of networks, layers, units, activation functions, error criterions, training procedures, etc. can be easily added, and is optimized for real-time applications. The whole code was compiled as a C++ library in order to be easily used in other projects (see next chapters of this thesis), and we developed a graphical user interface, called DeepSoft, in order to easily try different sets of parameters while continuously monitoring the training process.

The DeepSoft software and library first allow to load synchronous input and output data (**Fig. 22**).

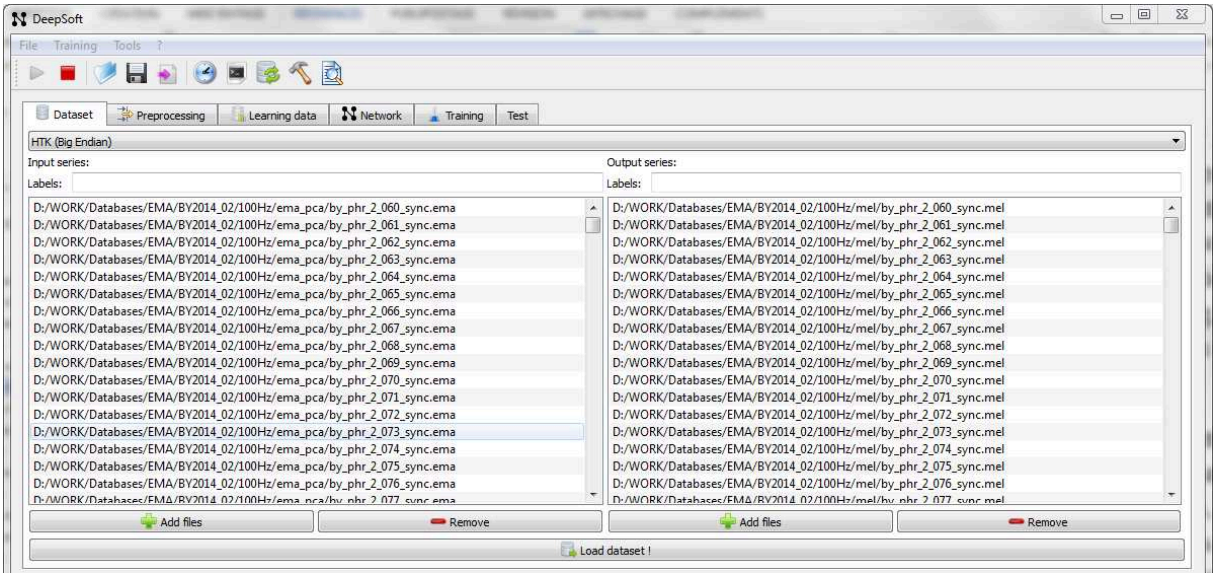


Fig. 22: DeepSoft – screenshot of the *"Dataset"* panel. Synchronous input (left) and ouput (right) data can be loaded from different file formats.

These input and output data can then be pre-processed by chaining several processing blocs, for instance select features, add temporal context to the data or remove silences (**Fig. 23**). The library can be easily extended to support new types of pre-processing.

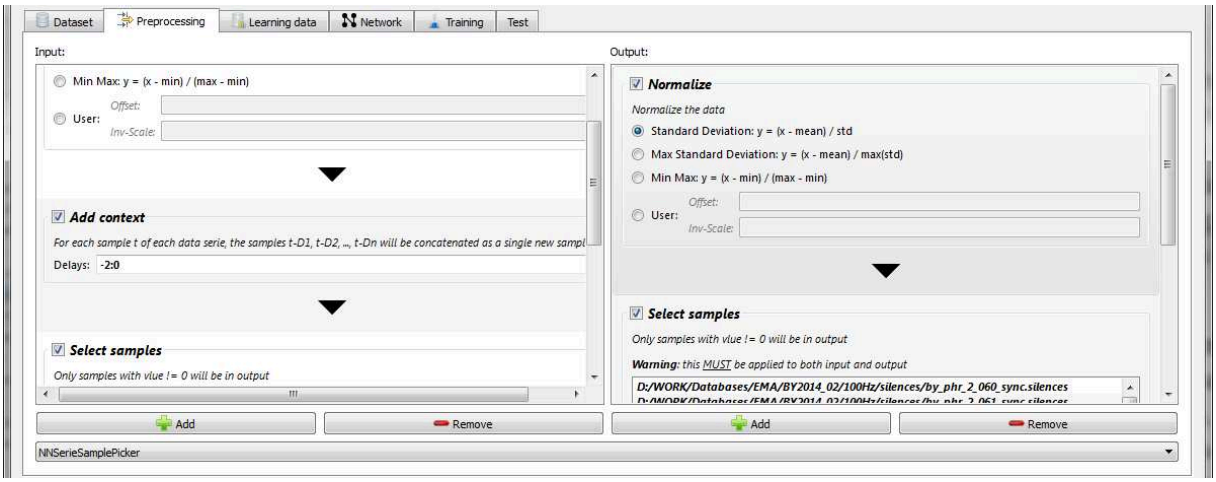


Fig. 23: DeepSoft – screenshot of the *"Preprocessing"* panel. Input (left) and output (right) data can be pre-processed by chaining different pre-processing blocs. For instance here, the output is first z-scored (top-*right bloc titled "normalize").*

The resulting dataset can be then partitioned into training, validation and test sets, for instance to perform cross-validation (**Fig. 24**).

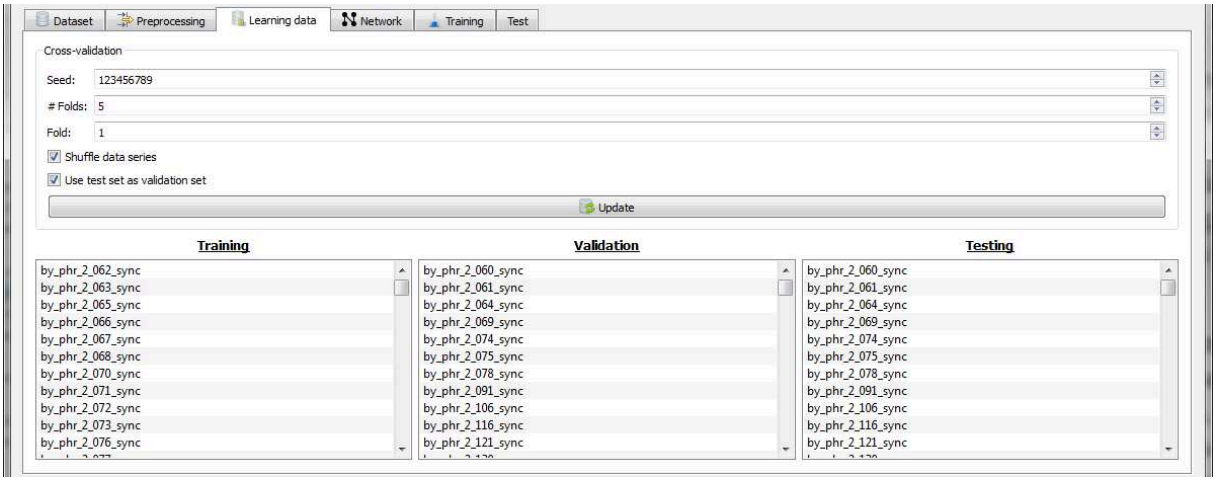


Fig. 24: DeepSoft – ***screenshot of the "Learning data" panel.*** This panel allows to split the data into training, test and validation tests, for instance to perform cros-validation.

The neural network is then designed by stacking different type of layers, for instance linear layers than connect two point-wise function layers (**Fig. 25**). Each layer is fully configurable – for instance for choosing the number of units and their activation function, and the library allows to easily add new types of layer in order to test different types of neural networks.

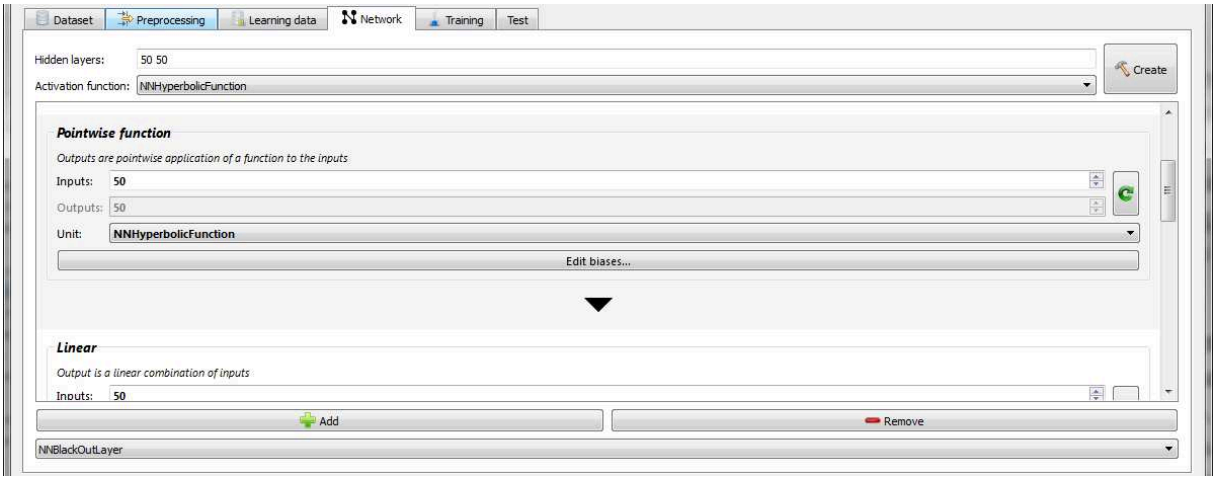


Fig. 25: DeepSoft – ***screenshot of the "Network" panel.*** The neural network is built by stacking layers. Here you can see that a linear layer is stacked over a point-wise function layer consisting of 50 hyperbolic units.

Once the network architecture is defined, the training panel (**Fig. 26**) allows to choose different training algorithms and their parameters, for instance conjugate gradient, as well as regularization methods, such as early-stopping or L2-regularization, and different choices of objective functions, for instance the mean-squared error. As for pre-processing and network layers, the library was developed so that adding new training algorithms, regularization methods or objective function can be easily done.

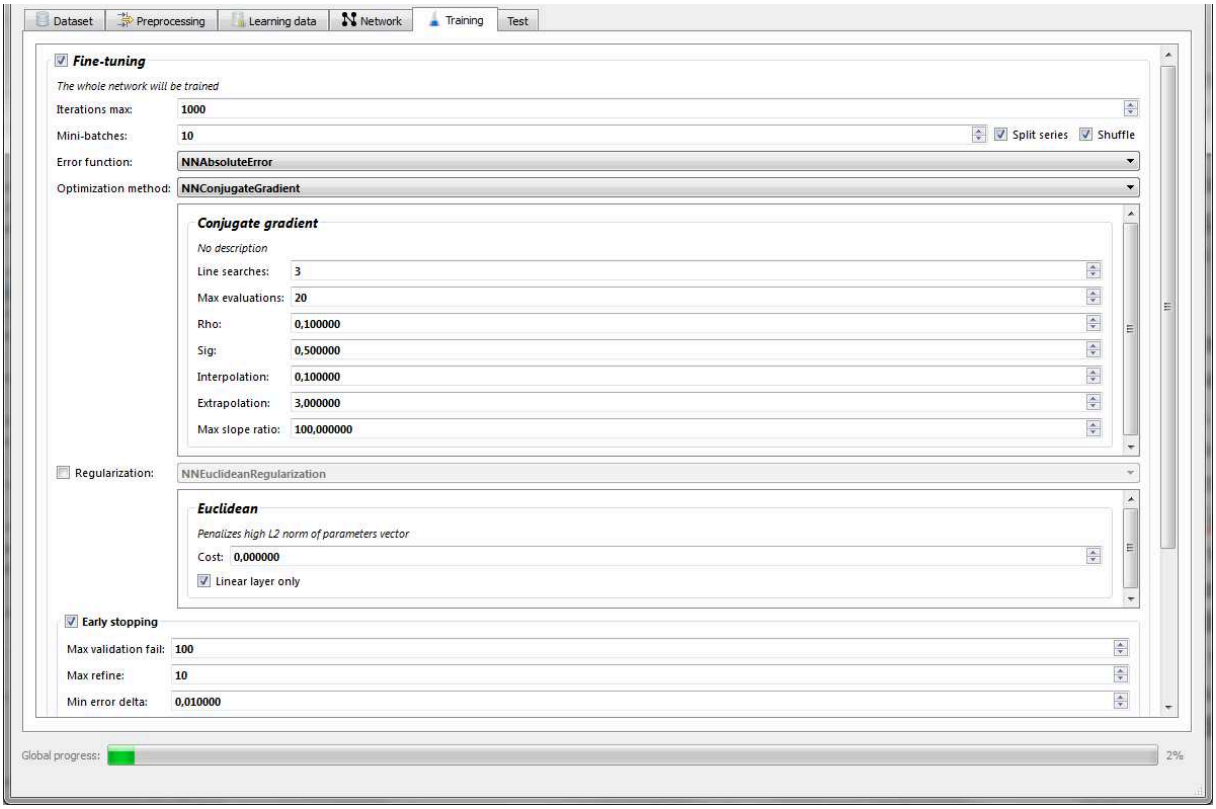


Fig. 26: DeepSoft – *screenshot of the "Training"* panel. This panel allows to choose and configure the training algorithm, including regularization methods and criterion function.

Once the training parameters are set, the network training can be performed while monitoring its progress, principally by monitoring the prediction error on the training and validation sets (**Fig. 27**), and previewing the prediction quality on the test set (**Fig. 28**). This real-time monitoring allows to test various training parameters and network architectures, while rapidly eliminating those for which the training is not satisfying, instead of waiting for the whole training to finish.

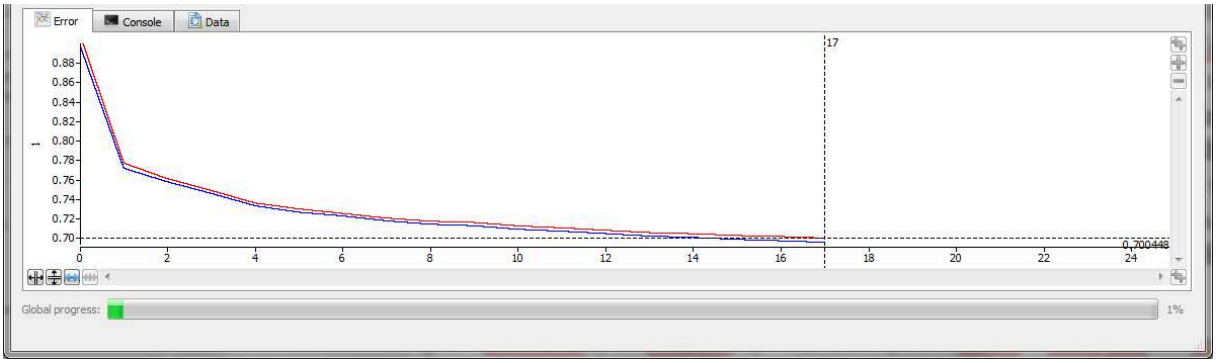


Fig. 27: DeepSoft – *screenshot of the "Error" panel.* At each epoch, the graph displays the error on the training set (blue line) and validation set (red line).

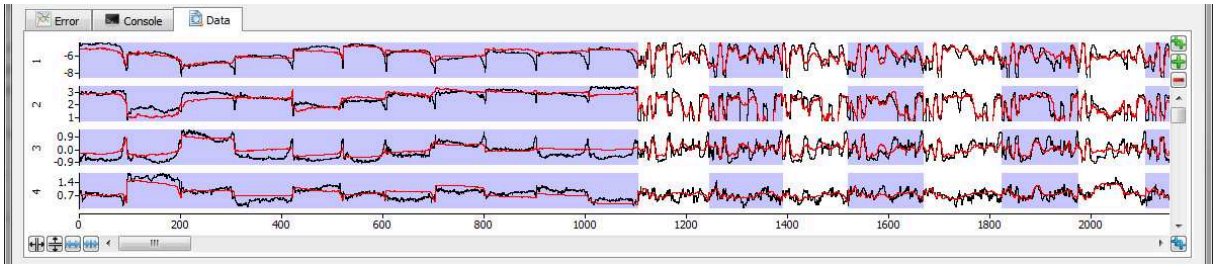


Fig. 28: DeepSoft – *screenshot of the "Data"* panel. This pannel allows to preview the network prediction (red line) compared to the ground truth (black line), for each item of the test set (each item is delimited by an alternating white/blue background).

Once trained, the neural network can be saved, as well as the pre-processing parameters, in a flexible and human-readable format that can be loaded and used through the C++ library from another software.

3. Articulatory-to-acoustic mapping and speech synthesis

GMMs and DNNs were used to build a mapping model allowing to transform articulatory data into acoustic data. Once trained, such mapping can be used to infer sequences of acoustic features (the MEL coefficients), from new articulatory data (the EMA signals). These estimated MEL coefficients define the MLSA filter which is then excited with a source signal representing the glottal activity in order to reconstruct the corresponding audio signal (see **Chapter 3**). **Fig. 29** summarize the synthesis chain when using the DNN-based mapping.

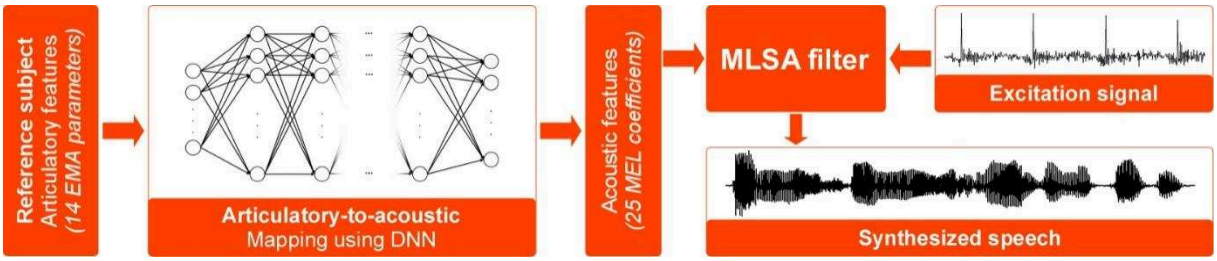


Fig. 29: Synthesis when using a DNN-based mapping. Using a DNN, articulatory features are mapped to acoustic features, which are then converted into an audible signal using the MLSA filter and an excitation signal.

a. Synthesis for the PB2007 corpus

For the PB2007 corpus, the excitation signal was designed as a white noise, mimicking the case of unvoiced sounds, as in whispered speech. Indeed, no information about the glottal activity is present in the EMA data, and whispered speech is mostly intelligible: in French, the following 6 pairs of phones are mostly discriminated by their voicing feature (the first phone of each pair being unvoiced): {/p/, /b/}, {/t/, /d/}, {/k/, /g/}, {/f/, /v/}, {/s/, /z/}, and {/ʃ/, /ʒ/}.

The sounds for the PB2007 corpus were synthesized both using the GMM- and the DNN-based mappings, with variable numbers of mixture components in the case of GMMs, and numbers of layers and units in the case of DNNs. Both input (articulatory) and output (acoustic) data were z-scored (subtraction of the mean and then division by the standard deviation) before being fed to the network, and data frames corresponding to silence periods were removed. To take into account the dynamic properties of speech, the GMM mapping used the first derivatives of the articulatory data (computed using finite differences with the previous and next frames),

while we concatenated each articulatory frame with its 2 preceding frames (30-ms time window context) for the DNN mapping. The GMM and DNN thus mapped the articulatory input features to 20 output mel-cepstrum coefficients, which were then converted into an audible speech signal using the MLSA filter and the excitation signal. Different GMMs and DNNs were tested on the PB2007 corpus in order to investigate best strategies for the articulatory-to-acoustic mapping.

b. Synthesis for the BY2014 corpus

As for the PB2007 corpus, the EMA did not contain any information on the glottal activity. To synthesize speech, we used two different excitation signals. The first excitation signal was designed so that all the synthesized sounds were voiced with a constant pitch, as opposed to the PB2007 synthesis for which we chose to have all the synthesized sounds unvoiced. That choice was made after preliminary experiments in which subjects reported the always-voiced synthesis to be more pleasant than the always-unvocied one. We thus designed an artificial template-based excitation signal using the glottal activity from a single vowel /a/. While such glottal activity could be recorded using an electroglottograph as in (Grimaldi and Fivela, 2008), here we estimated it using inverse filtering (Markel and Gray, 1976) (using the SPTK mlsadf tool). In short, inverse filtering works by extracting the mel coefficients from the original audio, then use these coefficient to compute the inverse filter of the MLSA. This inverse filter is then applied to the original audio in order to obtain the source excitation signal (**Fig. 30**). This signal was then used as a template to create an excitation signal by simply looping it. It was extracted from the isolated steady vowel /a/ in order to ensure an almost constant pitch (i.e. an almost period glottal activity) so that it could be looped.

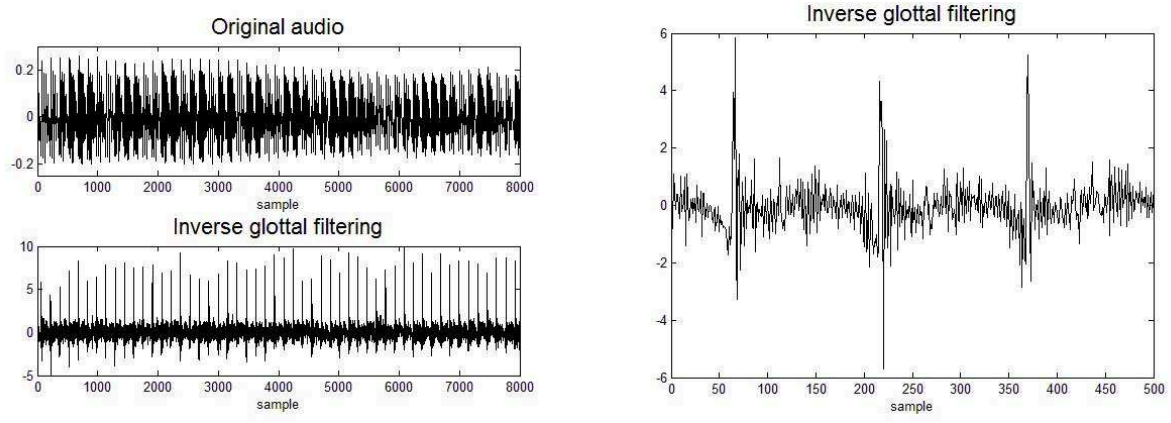


Fig. 30: Inverse glottal filtering. The top-left pannel represents the original audio signal extracted from an occurrence of the phone /a/. The bottom-left pannel represents the corresponding signal obtained by inverse filtering. The obtained signal is close to a pulse train which period corresponds to the pitch of the original /a/ signal. The left-pannel is a close-up on the first samples of the signal obtained by inverse filtering.

In order to evaluate the intelligibility loss when not using glottal activity estimation procedure (here by voicing all sounds), we performed as well the synthesis using another excitation signal. First the pitch was extracted from all the original audio signals using the SPTK pitch tool. Note that while we could have directly extracted the fundamental frequency from the glottal activity recorded through the electroglottograph, we chose here to extract the pitch from the audio, which is almost equivalent. That pitch was then used to generate an artificial excitation that was a white noise for unvoiced sounds, and a pulse train which period depended on the pitch value for voiced sounds, as can be seen in **Fig. 31**. The amplitude of each pulse is

equal to the square root of the pitch period, so that the power of the excitation signal is almost constant, in order to avoid that higher pitch sounds have more power.

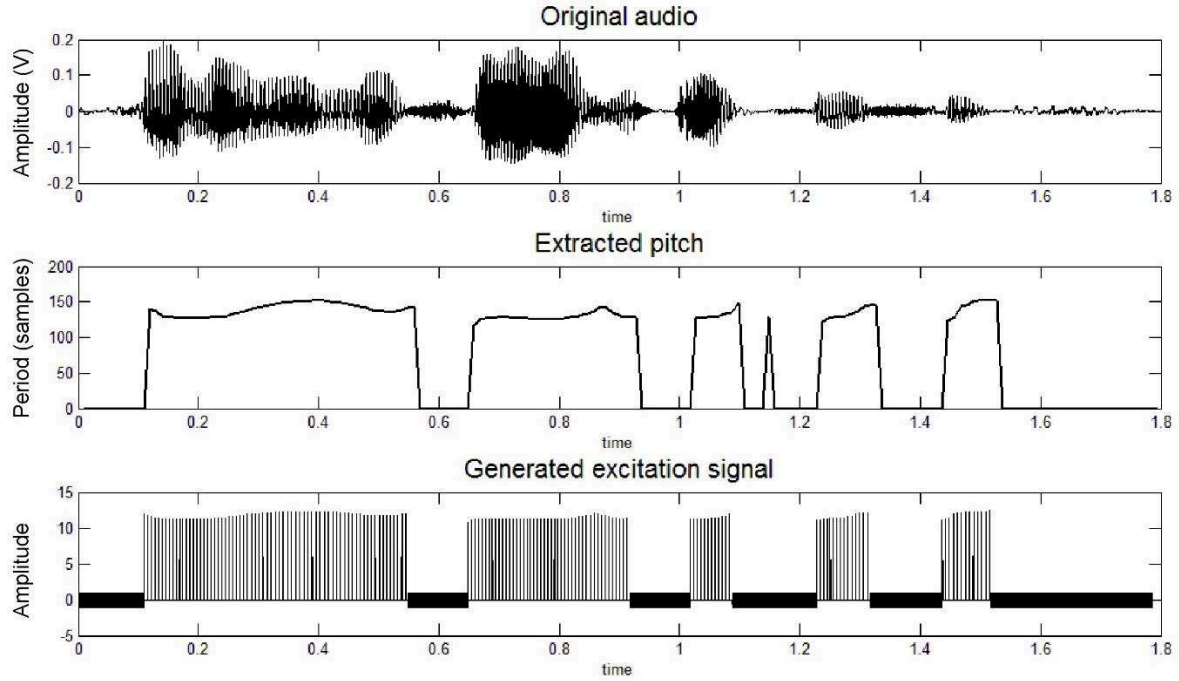


Fig. 31: Excitation signal generation. First the pitch is extracted from the original audio. This pitch is then used to generate an excitation signal which is white noise when the pitch is null (i.e. sounds are unvoiced) and a pulse train at the pitch period otherwise (i.e. sounds are voiced).

All the sounds for the BY2014 corpus were synthesized using only deep neural networks since results on the PB2007 corpus exhibited a superior robustness compared to the trajectory GMM approach. Following the results obtained from the PB2007 corpus, the DNN architecture was fixed, and had 3 hidden layers of 200 rectified linear units each. Both input (articulatory) and output (acoustic) data were z-scored (subtraction of the mean and then division by the standard deviation) before being fed to the network, and data frames corresponding to silence periods were removed. To take into account the dynamic properties of speech, we concatenated each articulatory frame with its 4 preceding frames (50-ms time window context compliant with a real-time implementation). The DNN thus mapped the articulatory input features to 25 output mel-cepstrum coefficients, which were then converted into an audible speech signal using the MLSA filter and the excitation signal.

III. Artificial degradation of the articulatory data

The intended application of the articulatory-based speech synthesizer is to build a speech BCI, where brain signals control in real-time the synthesizer. In this work, we considered speech synthesis from articulatory data since recent studies suggested that the speech motor cortex exhibited neural activity correlated to articulatory features during speech production, rather than acoustic features (Bouchard et al., 2013; Cheung et al., 2016). Thus, such speech synthesizer would be controlled by decoding the neural activity from the speech motor cortex. However, decoding of neural data often results in non-perfect signals, reflecting uncontrolled

fluctuations of brain activity or decoding and performance fluctuations. Thus, the ideal speech synthesizer must be robust to noisy articulatory inputs (Bocquelet et al., 2014). Moreover, the number of degrees of freedom (DoFs) that can be controlled by BCI remains limited and of the order of ten (Collinger et al., 2013; Ifft et al., 2013; Wodlinger et al., 2014). We thus evaluated, using the PB2007 corpus, the extent to which GMMs and DNNs mapping were robust to noise added on articulatory data and the minimum number of DoFs that were necessary to achieve intelligible speech. Different numbers of DoFs were also tested using the BY2014 corpus. In order to assess these properties, the articulatory data was artificialy degraded.

1. Noisy data

We tested the robustness of the articulatory-to-acoustic mapping by adding artificial noise to the test input articulatory data (no noise was added during the training step). We added white noise low-pass filtered below 20 Hz (as were the original EMA data) and re-centered. We tested different signal to noise ratio (SNR) values as defined by the ratio of the peak-to-peak amplitude of each articulatory signal by the standard deviation of the filtered noise. The noise amplitude was adjusted across EMA signals so that all had identical SNR.

2. Dimensionality reduction

In practice, accurate real-time BCI control of effectors can only be expected with a few degrees of freedom, typically less than 10. Hence, we tested to which extend it is possible to reduce the number of articulatory parameters, starting from all the parameters of our EMA database (12 for PB2007, 27 for BY2014), while preserving acceptable speech synthesis quality. We compared two main dimension reduction methods: the principal component analysis (PCA) and deep auto-encoders (DAE).

For the PB2007 corpus, we tested the performance of DNN- and GMM-based speech synthesis with all possible reduced dimensions, from 1 to 12, by feeding the originally trained models with reduced-then-recovered articulatory parameters, both using PCA and DAE.

For the BY2014 corpus, we only assessed the synthesis quality by reducing the articulatory parameters from 27, to 14, 10 and 7, using PCA only. While DAE gave very promising results the choice for PCA was motivated by its lack of hyper-parameters, as opposed to DAE (which are DNNs and thus have similar hyper-parameters).

Note that in all cases, the dimensionality reduction models were computed on the training data only.

a. Principal Component Analysis

Principal component analysis (PCA) is a well-known and widely used orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The principal components are sorted by variance, which means that the first component accounts for most of the variability in the original data, then the second is chosen so that it accounts for the highest

variance possible under the constraint that it is orthogonal to the first one, and so on. The resulting basis vectors are thus orthogonal and uncorrelated. By only keeping the first n components, PCA can be used to project the original data in a lower dimensionality space while preserving most of the original data variance. This projection can be inversed – with loss – in order to recover original data from reduced data.

Here PCA was applied to the input articulatory features in order to assess the effect of the number of articulatory parameters on the synthesis intelligibility. For the PB2007 corpus, PCA was applied keeping from 12 (i.e. all) to 1 components. For the BY2014 corpus, PCA was applied keeping 27 (i.e. all), 14, 10 and 7 components.

b. Deep Auto-Encoder

Deep auto-encoders (DAEs) are deep neural networks trained to reproduce their input as their output (Hinton and Salakhutdinov, 2006). Their architecture is symmetric with a "bottle-neck" linear middle layer containing fewer units than the input layer thus forcing the network to learn a dimensionality reduction of the input data (**Fig. 32**). Such a network can be spliced into two sub-networks: the encoding network, which reduces its input data X into $X_{reduced}$, and the decoding network, which allows recovering the data $\tilde{X}$ – with some loss - from the reduced data $X_{reduced}$.

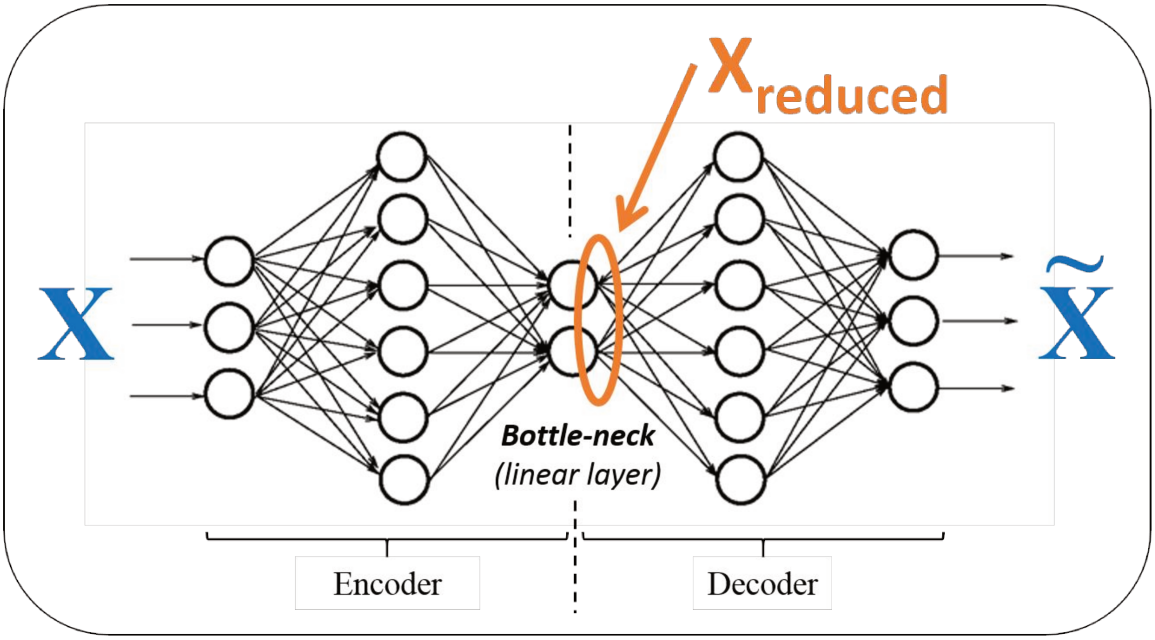


Fig. 32: Deep auto-encoder. A deep auto-encoder (DAE) is a symmetric deep ne*ural network with a "bottle-neck" that is* trained to reproduce its input as output. The purpose of the bottle-neck is to force the network to learn a representation of the original data using less parameters, and thus reducing the dimensionality of the original data. A trained DAE can then be split into the encoding part (left) which reduces the data, and the decoding part (right) which recovers original data from reduced data.

Here DAE were applied to the input articulatory features in order to assess the effect of the number of articulatory parameters on the synthesis intelligibility. It was applied only on the PB2007 corpus, with from 12 (i.e. all) to 1 reduced parameters. The encoder part of the DAE was first used to reduce all the articulatory data. Then, the decoder part of the DAE was added

in front of the articulatory-to-acoustic mapping model (GMM or DNN), in order to recover full articulatory parameters from the reduced ones (**Fig. 33**).

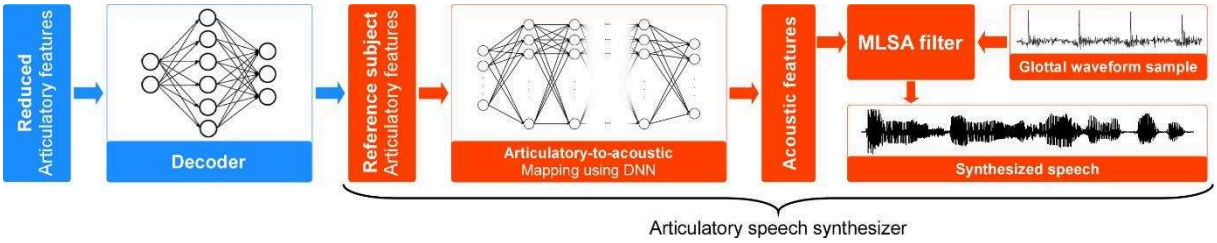


Fig. 33: Synthesis when using reducted articulatory data. The decoder part of the deep autoencoder (in blue) is added in front of the articulatory speech synthesizer (in orange) in order to recover full articulatory parameters from reduced parameters.

The DAEs were trained using the dimensionality reduction toolbox for Matlab (Maaten, n.d.). After preliminary experiments, the DAEs were chosen to have four hidden layers with all 50 units, except the middle one used for reduction, which number of units was equal to the desired number of reduced parameters.

IV. Evaluation of the speech synthesis intelligibility

Automatically evaluating speech synthesis results is not an easy task, so that most of the times a subjective evaluation performed by human subjects is needed. However, for fast prototyping, such as adjusting the training parameters, the number of mixture components, etc. an objective evaluation measure is needed. Indeed, subjective evaluation has two major limitations. First, evaluating sounds is a demanding task which tires rapidly the test subjects, so that they can only stay focused for about one hour. Second, when evaluating full sentences, subjects are biased once they have heard a sentence and will better recognize it if presented twice to them, even if the second occurrence is barely intelligible, improving the overall results and preventing any comparison. However, subjective evaluation remains the most reliable evaluation for speech. We thus used objective evaluation methods for preliminary analysis, which were then confirmed using restricted subjective evaluations.

1. Objective evaluation based on automatic speech recognition

One widely used measure is the mean mel-cepstral distortion (MCD) which compares two synthesized sounds S_1 and S_2 of identical length (Kubichek, 1993). The MCD measures the distance between the two sequences $\underline{\mathbf{m}^{(1)}}$ and $\underline{\mathbf{m}^{(2)}}$ of M mel-cepstrum coefficients, each sequence being of length N, respectively extracted from S_1 and S_2:

$$MCD(\boldsymbol{S1}, \boldsymbol{S2}) = \frac{10}{N.\ln(10)} \sum_{i=1}^{N} \sqrt{2 \sum_{j=1}^{M} \left(m_{i,j}^{(1)} - m_{i,j}^{(2)} \right)^2} \qquad \text{Eq. 26}$$

Note that Eq. 26 can be simplified to the mean root square error (MRSE) when removing the constant scaling factors:

$$MRSE(\boldsymbol{S1},\boldsymbol{S2}) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{M}\left(m_{i,j}^{(1)} - m_{i,j}^{(2)}\right)^2} \quad \text{Eq. 27}$$

It is important to mention that in the present work we are particularly interested in synthesis intelligibility – the ability to synthesize comprehensible speech, rather than in synthesis quality – the ability to synthesis pleasant speech similar to natural speech. Of course, quality takes some part in intelligibility, but the MCD is in fact more a measure of quality, rather than intelligibility: it actually measures how close two sounds are in term of spectral information, while in fact two /a/ pronounced by two different persons could have significantly different spectral content, and thus a high MCD, while both being equally comprehensible as being an /a/. This is why we chose to introduce another automatic evaluation method based on automatic speech recognition, which might be better correlated with the human perception of speech intelligibility.

Automatic speech recognition (ASR) is an entire research field and cannot be discussed in details in this manuscript since it is not the main focus of the thesis. The measure we will define here uses an ASR method to evaluate the phonetic content of a synthesized audio, in order to evaluate its intelligibility, and not its quality.

One method that has proven to be efficient for automatic speech recognition is based on Hidden Markov Models (HMMs). HMMs model the statistical relationship between some hidden states – phonetic units in the case of ASR – and some observable variables – in the case of ASR the speech spectrum (parametrized by mel-cepstrum coefficients). Once properly trained, an HMM estimates the most probable sequence of hidden states – the sequence of phonetic units – that have generated the given observable variables – the speech spectrum represented as a sequence of mel-cepstrum coefficients.

Here we used an HMM-based phonetic decoder trained on the spectral data of all the original database (thus one HMM for PB2007, and one HMM for BY2014) using a standard training procedure of context-dependent triphone tied-state HMM (Gales and Young, 2007). The recognition accuracy given by this decoder was used as a measurement of the quality of the synthetic spectral trajectories at the phonetic level. Such recognition accuracy is computed as so: the original sequence of phonemes is known from the training data, and the synthesized sequence of phonemes is obtained from the decoder. Since these sequences can be of different length (for instance the synthesis could have missed a word), they are first aligned using a string-alignment procedure based on dynamic programming (Young et al., 2009): the two sequences are aligned by minimizing the operations needed on the first phonemes sequence to transform it into the second one (**Fig. 34**). The possible operations on the phonemes are: none (the two phonemes are identical), substitution (replace a phoneme by another one), deletion (the phoneme is not present in the other sequence and must thus be deleted) and insertion (a phoneme was present in the other sequence but is missing in this one and must thus be inserted). Each operation is associated with a cost (usually, 0 for no operation, and 1 for all others), so that the algorithm finds the alignment that minimize the total cost of all operations. The accuracy is then defined as $Acc\% = 100\frac{(N-D-S-I)}{N}$, where N is the total number of phones, and S, D and I are respectively the number of substitutions, deletions and insertions.

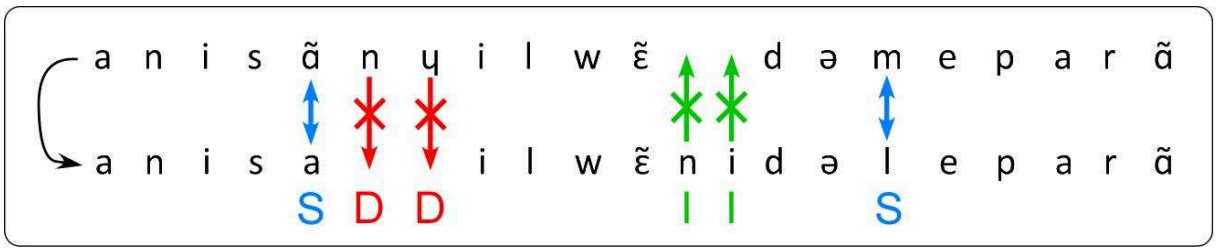


Fig. 34: Phoneme sequences alignment. Example of two aligned sequences of phonemes. S denotes a substitution, D a deletion and I an insertion.

The HMMs training and decoding were achieved using the Hidden Markov Model Toolkit (HTK, http://htk.eng.cam.ac.uk/). Here, neither dictionary nor language model was used for the recognition (as opposed to what is generally done in ASR), so that the HMM could not rely on language specific or context dependent information and only evaluated the phonetic content of the synthesized sounds.

This objective evaluation method was only used on the PB2007 corpus, mainly to evaluate the effect of hyper-parameters, such as the number of components in the Gaussian mixture, or the number of layers and units per layer in the deep neural network, as well as the effects of artificially degrading the articulatory data by adding noise to it or reducing its dimensionality. It was always applied to the test set (i.e. speech synthesized from articulatory data that were not part of the training set), which included various items, ranging from pseudo-words to short sentences.

2. Subjective evaluation using listening tests

While automatic evaluation gives indications on the synthesis intelligibility, the best way to assess it remains to perform perceptive evaluation by human subject. In order to avoid the influence of the linguistic context (vocabulary and grammar constraints from the French language), the synthesis intelligibility was mostly evaluated on isolated vowels and vowel-consonant-vowel (VCV) sequences, such as “apa”.

a. Evaluation on the PB2007 corpus

For the PB2007 synthesis, 11 subjects participated to an intelligibility test. All participants were French native speakers with no hearing impairment. The presented stimuli consisted of 10 French vowels /a/, /i/, /u/, /o/, /œ/, /e/, /y/, /ɑ̃/, /ɛ̃/, /ɔ̃/, and 30 VCV pseudo words made of the 10 consonants /p/, /t/, /k/, /f/, /s/ /ʃ/, /m/, /n/, /ʁ/, /l/, in /a/, /i/, /u/ contexts (e.g. “oto”). Since the synthesis on the PB2007 corpus was done using a white noise excitation, thus resulting in whispered speech with all sounds unvoiced, the 6 voiced consonants /b/, /d/, /g/, /v/, /z/, and /ʒ/ were excluded (see “Synthesis for the PB2007 corpus”). Participants were seated in quiet environment and instructed that they would be listening to isolated vowels or VCV sequences. For each utterance, they had to pick the corresponding vowel in the case of an isolated vowel, or the middle consonant in the case of a VCV sequence (forced choice paradigm). They were told that some of the sounds were noisy and difficult to identify, and thus to not evaluate the sound quality but only its intelligibility. Subjects could replay the stimuli as many times as necessary. This information was logged and in practice subjets only listened to each sound from

one to three times. No performance feedback was provided during the test. The recognition accuracy was defined as $Acc\% = 100\frac{R}{N}$ with R the number of correct answers for the N presented sounds of the test.

The seven following synthesis conditions were evaluated: analysis-synthesis ("anasynth"), GMM-based synthesis with noise (SNR = 10.0) and without noise, DNN-based synthesis with and without noise (SNR = 10.0), and DNN-based synthesis with and without reduced parameters (7 articulatory parameters obtained with a deep auto-encoder). Analysis-synthesis was performed by converting the audio signals into mel-cepstrum (MEL) coefficients, which were then directly converted back into audio signals using the MLSA filter and a white noise excitation. This conversion is not lossless, though it represents what would be the best achievable quality for the synthetic speech signal in the present context for the chosen acoustic parameterization.

b. Evaluation on the BY2014 corpus

For the BY2014 synthesis, 12 subjects participated to an intelligibility test similar to that of PB2007 (note that it was a different set of subjects). All participants were French native speakers with no hearing impairment. The evaluated stimuli consisted of the 10 vowels /a/, /i/, /u/, /o/, /œ/, /e/, /y/, /ã/, /ɛ̃/, and /ɔ̃/, and the 48 VCVs made of /p/, /t/, /k/, /f/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /z/, /ʒ/, /m/, /n/, /r/, and /l/, in /a/, /i/ and /u/ contexts (i.e., 'apa', 'iti', 'uku', and so on).

Each stimulus was synthesized and thus evaluated in 5 different conditions: 4 times using a pulse train excitation generated using the pitch extracted from the original audio (see **Fig. 31**) for each different number of reduced articulatory parameters (PCA with 27, 14, 10 and 7 components), and one time using the artificial template-based excitation signal (corresponding to a constantly voiced sound) with all 27 articulatory parameters. In the following, these 5 conditions are respectively denoted as Pitch_27, Pitch_14, Pitch_10, Pitch_7 and FixedPitch_27. This allowed us to evaluate both the influence of the number of articulatory parameters on the intelligibility, and the effect of using or not using a realistic glottal activity (i.e. not a constant pitch). Indeed, this glottal activity could be obtained by decoding the neural activity in future BCI application. An additional evaluation was performed for the two conditions Pitch_27 and Pitch_14, which consisted in directly transcribing 30 sentences (see Annex 1 for the list of these sentences). For each listener, half of the sentences were randomly picked from the first condition and the other half from the other condition, ensuring that each listener never evaluated the same sentence twice, and that all sentences were evaluated in both conditions.

The sounds were all normalized using automatic gain control, and played in random order for each subject at the same sound level through Beyerdynamic DT-770 Pro 80 Ohms headphones, while the listener was seated in a quiet environment. No performance feedback was provided during the test.

For the VCVs and vowels evaluation, participants were instructed to select from a list what they thought was the corresponding vowel in the case of an isolated vowel, or the middle consonant in the case of a VCV sequence (forced choice paradigm). Graphical user interface buttons were randomly shuffled for each subject in order to avoid systematic default choice

(e.g., always choosing the left button when unable to identify a sound). The subjects were told that some of the sounds were difficult to identify, and thus to choose the closest sound among the offered possibilities. The recognition accuracy was defined as $Acc\% = 100\frac{R}{N}$ with R the number of correct answers for the N presented sounds of the test. Since each item had exactly the same number of repetitions, the chance level was estimated by $Acc_{chance}\% = \frac{100}{C}$, with C the number of different item categories. The chance level was thus 1/10=10% for vowels, and 1/16≈6% for VCVs.

For the sentences, the subjects were asked to transcribe directly the sentences they were listening to. Results were evaluated using the word accuracy $WAcc\% = 100\frac{N-S-D-I}{N}$ (with N the total number of words, S the number of word substitutions, D the number of deletions and I the number of insertions), which is a commonly used metric in the field of automatic speech recognition (see "Objective evaluation based on automatic speech recognition").

3. Statistical analysis

a. Statistical analysis for the PB2007 synthesis

A 5-fold cross-validation was employed for evaluation of each model, allowing to obtain mean and standard deviation (SD) for each evaluation. Since the folds used to train the DNN and the GMM were identical, significant differences between results were assessed using the non-parametric Quade test with Conover correction (Quade, 1979), using recognition accuracy by phone for the objective evaluation (35 scores per condition) and recognition accuracy by participants for the subjective evaluation (11 scores per condition).

b. Statistical analysis for the BY2014 synthesis

Several listeners had to identify the same synthesized items, resulting in a binary answer (wrong or right), for each item and each listener. Statistical analysis of these results was thus performed using mixed logistic regression. For the VCVs and vowels, the following model was used: Result ~ (Segment + Condition)² + (1 | Listener), where Result is the binary answer (equals 0 if the item was wrongly identified, otherwise 1), Segment has two levels corresponding to the type of item (vowel or VCV), Condition has five levels corresponding to the five different conditions (Pitch_27, Pitch_14, Pitch_10, Pitch_7 and FixedPitch_27), and Listener has 12 levels corresponding to each listener that participated in the listening test. Multiple comparisons were made using contrasts according to (Hothorn et al., 2008). For the sentences, a paired Student test was performed to compare the results. All the tests were made using the R software, and packages lme4, multcomp and lsmeans.

V. Results

1. Convergence of the proposed approach for training DNNs for regression

Preliminary experiments were conducted to test different training methods for deep neural networks. We compared the conventional training, which consists of training the whole network at once, with our approach, which consists in successively adding layers in a pre-training phase before fine-tuning the whole network, as described previously. **Fig. 35** shows an example of the evolution of the mean-squared error (MSE) on the training set during the training of a DNN with five hidden layers of 50 units each, for both approaches.

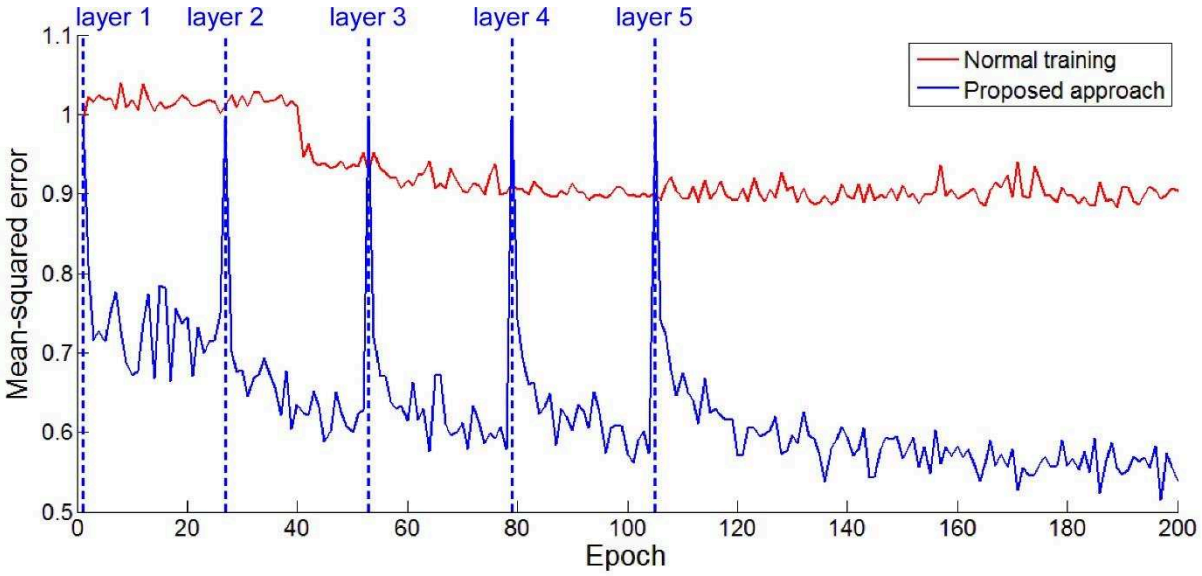


Fig. 35: Comparison of the conventional neural networks training with our approach. The graph shows the mean-squared error on the training set for each epoch, when using the conventional training (red line) and our proposed approach (blue line). Vertical dashed blue lines indicate the insertion of a new hidden layer during the pre-training stage.

In both cases, the initial MSE on the training set were identical (about 1). After 200 epochs, using the conventional training resulted in a MSE superior to 0.9, while using our approach resulted in a MSE inferior to 0.6. Note that this is not due to a superior number of epoch, since the network was trained for 200 epochs using the conventional training, and the same number of epochs was used for the proposed approach, including the number of epochs during pre-training (here 25 epochs were used for each pre-training step). As expected, each insertion of a new hidden layer resulted in a reset of the MSE since weights of the new layers were randomly initialized, but then the error rapidly decreased, speeding up the training process.

2. PB2007 corpus

a. Influence of GMM hyper-parameters

Objective evaluations were conducted for various numbers of mixture components in the GMM, from 16 to 256. As shown on **Fig. 36**, the recognition accuracy increased as the number of components increased, before stabilizing as long as at least 128 components were used. For

the following tests, we thus chose a GMM with 128 components, a choice that is consistent with previously published results (Toda et al., 2008). This was also motivated by the fact that fitting 256 components often led to ill-conditioned covariance matrices. This corresponded to an accuracy of 75.68 ± 1.27%.

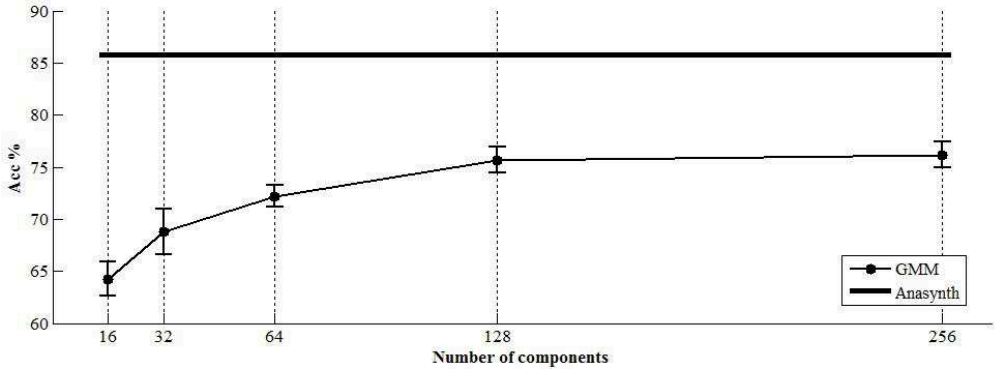


Fig. 36: Influence of the GMM hyper-parameters. The thin line shows the phone recognition according to the number of mixture components in the GMM-based mapping (mean±SD). The thick line represents the recognition accuracy on the anasynth audio.

b. Influence of DNN hyper-parameters

We also conducted objective evaluations of speech synthesis for various DNN architectures, with different numbers of layers (1 to 4) and numbers of units per hidden layer (20, 50, or 100) identical across hidden layers. As shown in **Fig. 37**, adding more units for a given layer increased recognition accuracy, while adding more layers first led to an increase before a stabilization or small degradation in accuracy. Overall, a good compromise was to use a DNN with 3 hidden layers of 100 units each, ensuring an accuracy of 71.13 ± 2.75%.

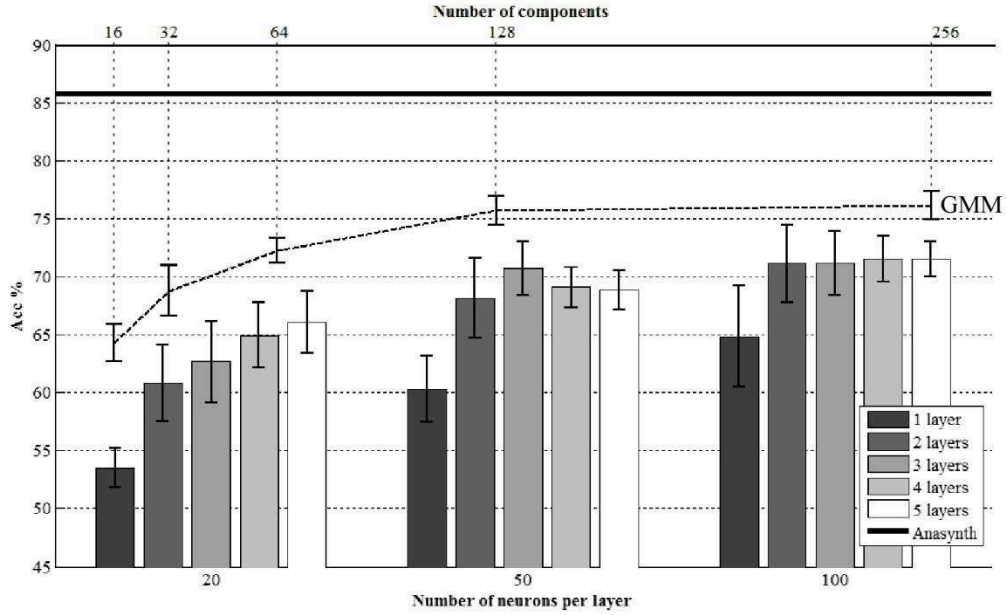


Fig. 37: Influence of the GMM and DNN hyper-parameters. The dashed line shows the phone recognition according to the number of mixture components in the GMM-based mapping (mean±SD). bar plot shows the phone recognition accuracy according to the number of layers and the number of units per layer using DNN-based mapping (mean±SD). The thick line represents the recognition accuracy on the anasynth audio.

Fig. 38 further shows the recognition accuracy by vowel and consonant as given by the subjective evaluation for this chosen architecture of 3 hidden layers of 100 units each.

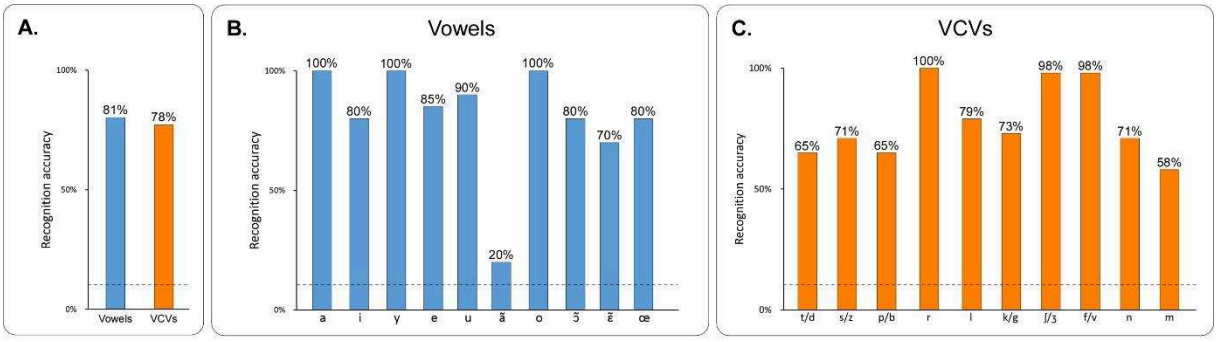


Fig. 38: Recognition accuracy by vowel and consonant for the PB2007 synthesis with a DNN of 3 hidden layers of 100 units each. A – Owerall recognition accuracy for vowel and VCVs. The dashed line indicates chance level. B – Recognition accuracy by isolated vowel, for the subjective evaluation. The dashed line indicates chance level. Vowels are sorted by number of occurences in the training set, from higher to lower. C – Recognition accuracy according to the middle consonants of the VCVs, for the subjective evaluation. The dashed line indicates chance level. Consonants are sorted by number of occurences in the training set, from higher to lower (for consonant pairs, the sum of occurences of each consonant in the pair was used).

For vowels, 4 out of 10 vowels had recognition accuracy above 90% (/o/, /u/, /a/ and /y/), while the worst recognition accuracy was achieved for the nasal vowels /ɑ̃/ (20%) and /ɛ̃/ (70%), which was expected since there was no velum information in the EMA data. For the consonants, 3 out of 10 groups of consonants had recognition accuracy above 90% (/r/, {/ʃ/, /ʒ/} and {/f/, /v/}), while the worst results were achieve for the nasal consonant /m/ (58%), and the pairs of plosive consonants {/p/, /b/} (65%) and {/t/, /d/} (65%).

In **Fig. 38**, vowels and consonants are sorted by number of occurences, from higher to lower (the sum of the occurences of each consonant was used for pairs). Note that there is no clear correlation with the number of occurrences of each phone in the training corpus, since for instance the corpus contained few instances of /f/ and /v/, and a larger number of /t/ and /d/.

c. Comparison of GMM and DNN

We then compared GMM-based and DNN-based synthesis using both the objective (HMM phonetic decoding) and the subjective (listening) tests. Consistent results were obtained, as shown in **Fig. 39**. In the objective test, GMM recognition accuracy reached 75.68% and the DNN, 71.13%. In the subjective test, the GMM recognition accuracy was 66.59% and 69.77% for the DNN. Both GMM and DNN recognition accuracies were below the recognition accuracy on original audio ($P < 10^{-4}$ for the objective evaluation and $P < 0.002/0.01$ for the subjective one for GMM/DNN), which was 85.77% for the objective evaluation, and 87.95% for the subjective one. The GMM performed slightly better than the DNN in the objective evaluation ($P < 10^{-3}$) while no significant difference was observed between both models in the subjective evaluation ($P > 0.3$). This might be caused by the fact that HMMs share similar properties with GMMs, as previously discussed.

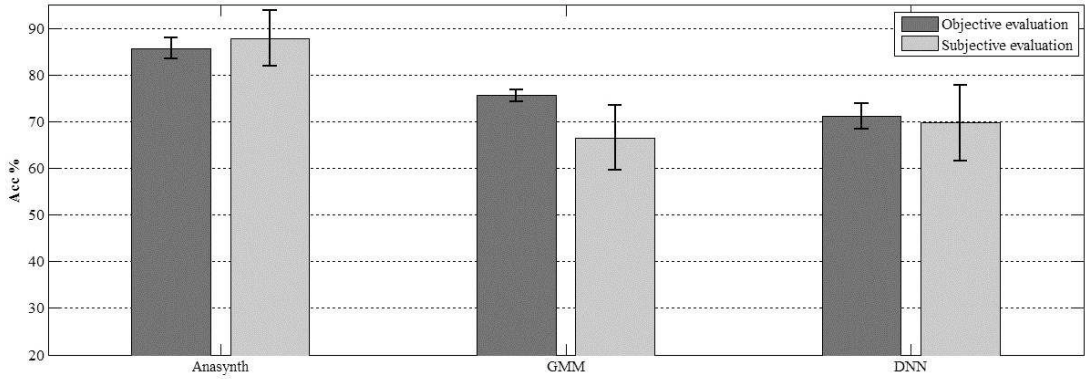


Fig. 39: Comparison of GMM and DNN mappings for both objective and subjective evaluations. Each bar corresponds to the recognition accuracy (mean±SD).

d. Speech synthesis from reduced articulatory data

Then, we evaluated the performance of the speech synthesizer when degrading articulatory inputs. We first reduced the dimensionality of the data. Either PCA or DAE were combined with the GMM-based and the DNN-based mappings to test successive dimension reduction from 12 to 1 articulatory parameter. For less than 9 reduced parameters, the use of DAE led to better results, both for the GMM- ($P < 10^{-4}$) and the DNN-based mappings ($P = 0.01$), while no significant difference was observed with 11 and 12 parameters (**Fig. 40**). Using 7 or more DAE-reduced parameters allowed obtaining a recognition accuracy of above 60% both for the GMM- and the DNN-based mappings, while 9 or more parameters were needed to achieve the same accuracy when using PCA. Moreover, no significant difference was observed between GMM- and DNN-based mappings for less than 9 reduced parameters obtained by PCA ($P > 0.8$), while the GMM results were slightly better than the DNN for more than 3 reduced parameters obtained by DAE ($P = 0.01$).

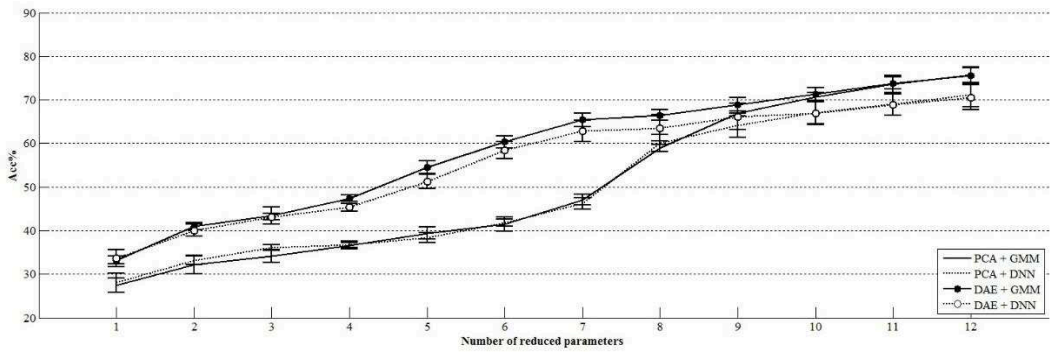


Fig. 40: Evaluation of the PB2007 synthesis with reduced articulatory data. Phone recognition accuracy (mean±SD) with reduced parameters obtained both by principal component analysis (PCA) and deep auto-encoders (DAE), and with both GMM- and DNN-based mappings.

e. Speech synthesis from noisy articulatory data

Both GMM- and DNN-based mappings were then objectively evaluated with noisy input data with different SNR (from 2 to 20, **Fig. 41**). With the GMM-based mapping, a recognition

accuracy above 60% was reached with a SNR of more than 20, while the DNN-based mapping obtained more than 60% recognition accuracy with a SNR higher than 10. The DNN-based mapping generally obtained better recognition accuracy than the GMM-based mapping ($P < 10^{-4}$).

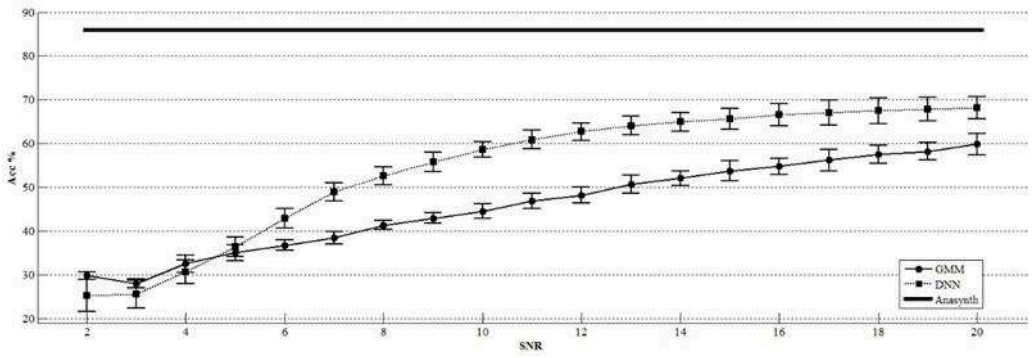


Fig. 41: Evaluation of the PB2007 synthesis on noisy articulatory data. Phone recognition accuracy (mean±SD) on noisy data as a function of the signal to noise ratio (SNR), both for GMM- and DNN-based mappings.

A subjective test was then conducted for GMM- and DNN-based mapping of noisy EMA data (SNR=10, which corresponds to 44.5% and 58.6% of recognition accuracy for the GMM- and the DNN-based mapping respectively, in the objective test). This test also included DNN-based mapping of DAE-reduced data (7 parameters) with and without noise addition (**Fig. 42**). The GMM-based mapping obtained a recognition accuracy of 32.3%, while the DNN-based mapping obtained 59.3% with no parameters reduction, and 53.9% when using 7 DAE-reduced parameters. Consistently with the objective evaluation, the DNN-based mapping was found to perform better than the GMM-based mapping in noisy condition ($P<10^{-4}$). Moreover, the DNN-based mapping with reduced and noisy parameters performed better than the GMM-based mapping with full and noisy parameters ($P = 0.01$). Finally, no significant difference in subjective accuracy of the DNN-based mapping with reduced parameters was observed between clean and noisy conditions ($P > 0.2$).

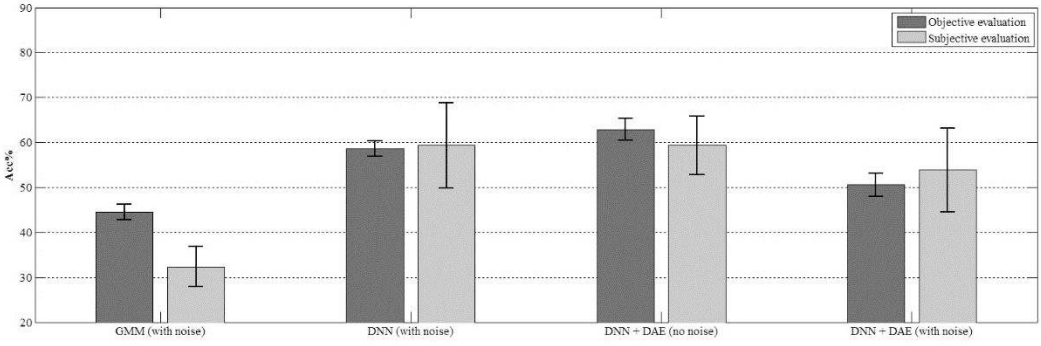


Fig. 42: Objective and subjective evaluation of the PB2007 synthesis with both noisy and reduced parameters. The bar plot shows the recognition accuracy (mean±SD) based on objective and subjective evaluations of GMM- and DNN-based mapping on noisy articulatory data (SNR = 10), and on reduced data (DAE with 7 reduced parameters) for the DNN-based mapping.

f. Conclusion on the PB2007 synthesis

We compared the perfomance of two articulatory-to-acoustic mapping methods, one using the state-of-the-art method which relies on trajectory Gaussian mixture models, and another one based on deep neural networks. Both mapping approaches were evaluated on clean and

noisy articulatory data, with and without reducing the dimensionality of the input articulatory parameters, in order to assess their robustness. The two methods where objectively evaluated using a HMM-based speech recognition method, and subjectively evaluated with a listening test. Objective and subjective evaluations were consistent and pointed out that the DNN-based mapping was reaching a phone recognition accuracy of around 70% which is almost similar to the results obtained with the GMM-based mapping. It also showed that DNNs were more robust to noise than GMMs. Results on the dimensionality reduction showed that DAEs were more appropriate than PCA, both for GMM- and DNN-based mappings. Finally, it is important to note that the DNN-based mapping has a very low computational cost once the network has been trained, and is thus compatible with real time applications such as BCIs. For these reasons, we chose to only use deep neural networks in the following experiments. Moreover, since both GMM- and DNN-based mapping showed results still far below the anasynth results, we chose to record a new and larger articulatory-acoustic corpus, the BY2014 corpus, in order to improve synthesis (in particular for nasals and plosives).

3. BY2014 corpus

a. Evaluation results on vowels and VCVs

Fig. 43 summarizes the result of the subjective listening test. The recognition accuracy was better for vowels than for consonants for FixedPitch_27, Pitch_27 and Pitch_7 ($P < 0.01$), while this difference was only a trend for Pitch_14 ($P = 0.0983$) and no difference was found for Pitch_10 ($P > 0.99$).

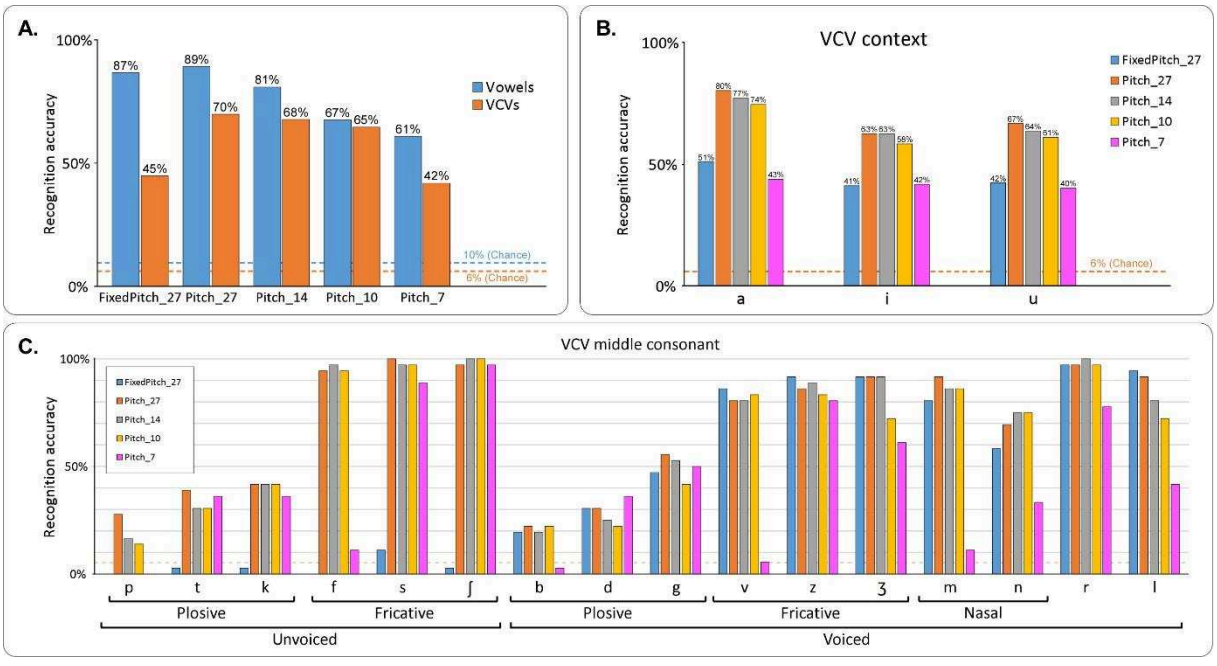


Fig. 43: Subjective evaluation of the intelligibility of the BY2014 speech synthesizer. A – Recognition accuracy for vowels and consonants for each of the 5 synthesis conditions. The dashed lines show the chance level for vowels (blue) and VCVs (orange). B – Word recognition accuracy for the sentences, in both conditions Pitch_27 and Pitch_14. C – Recognition accuracy of the VCVs regarding the vocalic context, for the 5 synthesis conditions. The dashed line shows the chance level.

For vowels, the recognition accuracy was far above chance (chance level = 10%) for all conditions ($P < 0.01$, **Fig. 43**-A) and decreasing when decreasing the number of articulatory parameters, ranging from 89% for Pitch_27 to 61% for Pitch_7. Taking Pitch_27 as reference, this decrease was not found significant for Pitch_14 ($P = 0.7116$), and significant for Pitch_10 and Pitch_7 ($P < 0.01$ in both cases). No statistically significant difference was observed when not using the glottal activity versus when using the glottal activity (FixedPitch_27 = 87%, Pitch_27 = 89%, $P > 0.99$).

For the consonants, the recognition accuracy was also far above chance (chance level = 6.25%) for all conditions ($P < 0.01$, **Fig. 43**-A). A decrease in recognition accuracy was also observed when decreasing the number of articulatory parameters, ranging from 70% for Pitch_27 to 42% for Pitch_7. However, taking Pitch_27 as reference, this decrease was not significant for Pitch 14 ($P > 0.99$) and Pitch 10 ($P = 0.6328$), and only significant for Pitch_7 ($P < 0.01$). A significant difference was observed when not using the glottal activity (FixedPitch_27 vs Pitch_27, $P < 0.01$). The differences in recognition accuracy for each condition were studied regarding the vowel of the VCV (**Fig. 43**-B) and the consonant (**Fig. 43**-C). Overall the intelligibility was higher when the consonant was in /a/ context (/a/ being the most represented phone in the corpus, see **Fig. 19**-A) than when in /i/ and /u/ context ($P < 0.01$), and no significant difference was observed between /i/ and /u/ contexts ($P > 0.99$): for instance, for Pitch_27, accuracy decreased from 80% for /a/ context, to 63% and 67% for /i/ and /u/ contexts respectively. Regarding consonants (**Fig. 43**-C), no clear differences were observed between the three synthesis Pitch_27, Pitch_14 and Pitch_10 except for /p/, /l/, /d/, /g/ and /ʒ/. Clear differences between these three conditions and Pitch_7 were observed for consonants /p/, /f/, /b/, /v/, /ʒ/, /m/, /n/, /r/ and /l/. Clear differences were also observed between FixedPitch_27 and Pitch_27 for the unvoiced consonants /p/, /t/, /k/, /f/, /s/, and /ʃ/. Conversely, no significant differences between FixedPitch_27 and Pitch_27 were found for all the voiced consonants, which includes all the consonants chosen for the real-time closed loop synthesis that does not use the glottal activity (i.e. it is similar to FixedPitch_27). All conditions taken together, best results (at least one condition above 90%) were achieved for the fricative consonants /f/, /s/, /ʃ/, /z/, and /ʒ/, the nasal consonants /m/ and /n/, and /l/. Worst results (all conditions below 50%) were achieved for the plosive consonants /p/, /t/, /k/, /b/ and /d/. Note that there is no clear correlation with the number of occurrences of each phone in the training corpus, since the corpus contained few instances of /ʃ/, and a large number of /t/ (**Fig. 19**-A).

Analysis of the confusion matrices can enlighten the sources of synthesis errors (**Fig. 44**). Each row i of a confusion matrix M corresponds to the ground truth phone p_i, while column j corresponds to the phone p_j recognized by the listeners, so that a diagonal value $M_{i,i}$ corresponds to the proportion of occurrences of the phone p_i that were correctly recognized, and a value $M_{i,j}$ outside the diagonal corresponds to the proportion of occurrences of p_i that were substituted by p_j. The order of the rows and columns of the confusion matrices were automatically sorted in order to emphasize the main confusions by forming high value blocks near the diagonal.

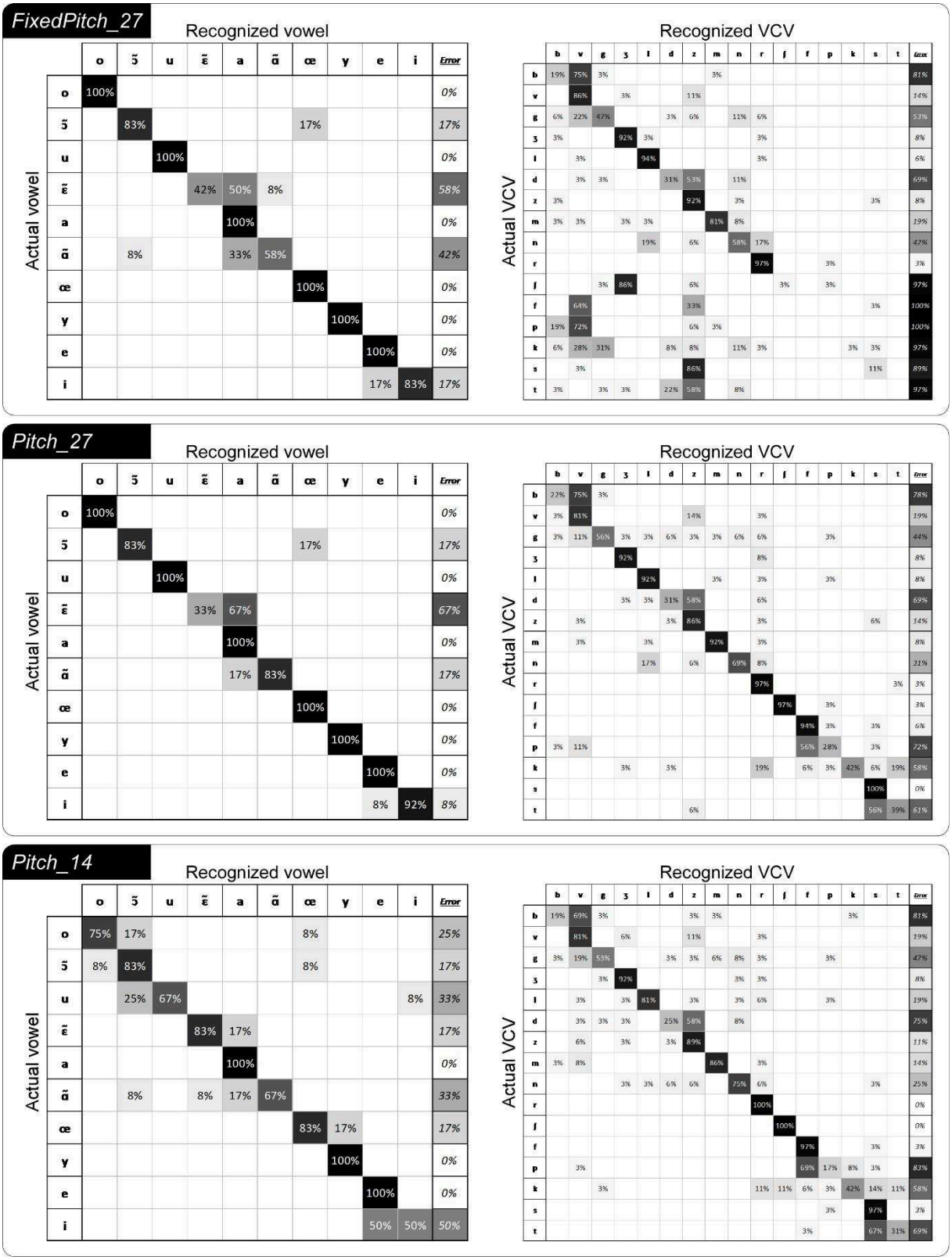


Fig. 44: Confusion matrices of the subjective evaluation of the intelligibility of the BY2014 speech synthesizer. Confusion matrices for vowels (left) and consonants (right), for each of the three conditions FixedPitch_27, Pitch_27 and Pitch_14. In the matrices, rows correspond to ground truth while columns correspond to user answer. The last column indicates the amount of errors made on each phone. Cells are colored by their values, while text color is for readability only.

The confusion matrices of the perceptual listening test for the condition Pitch_27 (**Fig. 44**, middle row) reflect the global good quality of this synthesis (indicated by the fact that they are

near-diagonal matrices). For vowels, six out of the ten vowels were always correctly recognized (/o/, /u/, /a/, /œ/, /y/ and /e/). Main errors come from confusions between /ɛ̃/ and /a/ (67% of /ɛ̃/ were recognized as /a/), and other errors come from confusions between /ɑ̃/ and /a/ (17% of /ɑ̃/ were recognized as /a/), and between /ɔ̃/ and /œ/ (17% of /ɔ̃/ were recognized as /œ/. For consonants, main confusions came from /b/ being recognized as /v/ (75%), /d/ being recognized as /z/ (58%), /p/ being recognized as /f/ (56%) and /d/ being recognized as /z/ (58%). Other more minor errors come from /g/ being recognized as /v/ (11%), and /k/ being recognized as /r/ (19%) and /t/ (19%).

By comparing confusion matrices of Pitch_27 with those of FixedPitch_27, we can observe that not using the glottal activity resulted in increased confusions mainly for the vowel /ɑ̃/ (accuracy going from 83% for Pitch_27 to 58% for FixedPitch_27) while no clear difference can be observed for the other vowels. Note that between the two conditions Pitch_27 and FixedPitch_27, the articulatory-to-acoustic model remains the same, the only change being the excitation signal that is used for the final synthesis with the MLSA filter. Importantly, for the consonants, not using the glottal activity resulted in a drastic decrease in the recognition accuracy of all the unvoiced consonants /p/, /t/, /k/, /f/, /s/ and /ʃ/, while all the voiced consonants remained recognized with similar accuracy. Indeed, /p/ was mainly recognized as /v/ (72%), /t/ as /z/ (58%), /f/ as /v/ (64%), /s/ as /z/ (86%), and /ʃ/ as /ʒ/ (86%). Note that /v/ is the voiced counterpart of /f/, /z/ of /s/ and /ʒ/ of /ʃ/. Hence, the use of the template-based excitation naturally leads to a predictable shift of the unvoiced consonants to their more or less corresponding (in terms of place of articulation) voiced counterparts.

By comparing the confusion matrices of Pitch_27 with those of Pitch_14, we can observe that there is no clear pattern of increased confusions. This confirms the results previously obtained from **Fig. 43**, where no significant differences between Pitch_27 and Pitch_14 were found for both vowels and consonants.

b. Evaluation results on full sentences

The results of the subjective evaluation on sentences are presented in **Fig. 45**. While the recognition accuracy for Pitch_27 and Pitch_14 was below 90% for vowels and below 70% for consonants, the word recognition accuracy for the sentences was above 90% for both conditions (96% for Pitch_27 and 92% for Pitch_14). The difference in recognition accuracy for Pitch_27 and for Pitch_14 is statistically significant (P = 0.015).

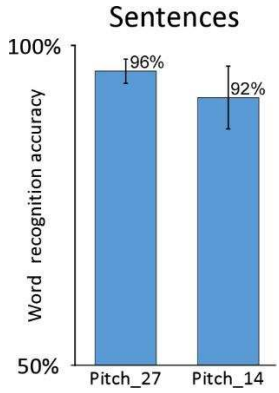


Fig. 45: Subjective evaluation of the intelligibility of the BY2014 speech synthesizer on sentences. Word recognition accuracy for the sentences, for both conditions Pitch_27 and Pitch_14.

c. Conclusion on the BY2014 synthesis

The analysis of the synthesis results was performed deeper with the BY2014 synthesis than with the PB2007 one. Several versions of the synthesizer were built to assess the effect of the number of articulatory parameters (27, 14, 10 and 7), and the effect of using or not using glottal activity (by comparing synthesis using a constant artificial pitch, and the original pitch). The phone recognition accuracy for offline reference synthesis was far above chance level for all five tested parameterizations, and fully intelligible speech sentences could be produced (**Fig. 45**). Most errors on vowels were made between vowels that had close articulatory positions (e.g. /i/ and /e/, see **Fig. 19**-B). Regarding consonants, most errors were made on the plosive consonants, and main confusions were observed within pairs of consonants corresponding to relatively similar articulatory movements in terms of place of articulation: for instance, /b/ is a labial consonant and /v/ is a labio-dental, and /d/ is a dental or an alveolar and /z/ is an alveolar. For /b/-/v/ confusion, this could be explained by a positioning of the EMA coils too far from the lip edges, resulting in a tracking of the lips by the EMA system that did not allow to capture sharp differences between /b/ and /v/ lip movements. A similar interpretation can be given for /d/-/z/ confusions, since in practice the coil had to be attached more than 5 mm back from the tongue tip (see **Fig. 18**-A). Moreover, results showed that the accuracy on the VCVs was correlated to the vocalic context, with consonant in /a/ context having a better recognition accuracy. This could be explained by the fact that the phone /a/ is more largely present in the training corpus than the phones /i/ and /u/ (see **Fig. 19**-A). However, this is not consistent with the fact that some phones that are less represented in the corpus, like /ʃ/, have high recognition accuracy, while other phones that are largely represented, like /d/, have low recognition accuracy. Another possible explanation is that /a/ is the most opened vowel and thus VCVs in /a/ context are performed by movements of higher amplitude, which could be more discriminant. By removing the glottal activity information (here by using a constant pitch), we found that the recognition accuracy was dramatically decreased for all unvoiced consonants, while remaining roughly the same for all voiced consonants and vowels (see **Fig. 43** and top and middle rows of **Fig. 44**). The unvoiced consonants were thus confused with their voiced counterparts (e.g. /ʃ/ with /ʒ/), or with the voiced counterpart of the consonant they were already confused with (for instance, /p/ was originally confused with /s/ in the Pitch_27 condition and was then confused with /ʒ/ when using a constant pitch, in the FixedPitch_27 condition).

Regarding the number of articulatory parameters, the results showed that using 14 articulatory parameters yields intelligibility scores that were close to the best scores achieved with 27 parameters. Interestingly, using 10 parameters did not significantly impact the intelligibility of consonants, but started to affect that of vowels, although the accuracy remained at the high level of 67%. Decreasing further the number of parameters down to 7, significantly impacted the intelligibility of both vowels and consonants. Finally, although the accuracy on consonants was inferior to 70% for 27 and 14 articulatory parameters, this was enough to produce very intelligible sentences, with word recognition accuracy superior to 90% (see **Fig. 45**). This can be explained by the fact that most confusions were made with similar consonants, thus ensuring a good intelligibility when constrained with closed vocabulary and syntactic rules. Thus, overall, the number of parameters required to achieve a sufficient intelligibility is of the order of 10, which is the number of degrees of freedoms that could be controlled successfully in recent state of the art BCI experiments (Wodlinger et al., 2014). It should be noted that the reduction in the number of parameters was done here in a drastic way either by dropping

parameters or by PCA, while more efficient dimensionality reduction techniques could be envisioned such as the autoencoders that we previously started to investigate on the PB2007 corpus.

VI. Conclusion on the articulatory-based speech synthesis

In this chapter we presented an approach to synthesize speech from articulatory movements using deep neural networks (DNNs), which was compared to a state-of-the-art approach using Gaussian mixture models (GMMs).

First, the synthesis was performed using a previously recorded articulatory-acoustic corpus, the PB2007 corpus. This corpus contained synchronous articulatory and acoustic data, which allowed to trained GMMs and DNNs in order to capture the relationship between articulatory and acoustic features. These models were then used to map new articulatory data to acoustic features, which were then converted into audible speech using the MLSA filter. Two approaches were used to evaluate the synthesis intelligibility: an objective one, based on automatic speech recognition, and a subjective one, through listening tests by human subjects. The objective evaluation suggested that the GMM-based synthesis was more intelligible than DNN-based synthesis. However, the subjective evaluation showed no significant difference: both approaches were able to synthesize the French phones with about 70% average accuracy. We then evaluated the synthesis intelligibility when reducing the dimensionality of the articulatory parameters. Two dimensionality reductions techniques were used: principal component analysis (PCA) and deep auto-encoders (DAE). Results suggested that DAEs were more appropriate than PCA, both for GMM- and DNN-based mappings (**Fig. 40**). However, it is important to mention here that DAE have many hyper-parameters (see "Choice of DNN hyper-parameters"), while PCA has none. In order to further compare the GMM- and the DNN-based synthesis, their robustness to noisy articulatory inputs was evaluated. Both objective and subjective evaluations showed that the DNN-based mapping is more robust to noisy articulatory inputs than the GMM-based mapping (**Fig. 41** and **Fig. 42**).

For these reasons, DNNs were preferred to GMMs for future experiment, which was as well motivated by the fact that DNNs are more compatible with real-time application than the trajectory GMM since they require less computational power once trained and work on a frame-by-frame basis while the GMM-based mapping is working on full sequences of articulatory data. In order to further improve the synthesis quality, the choice was made to record a new and larger corpus, the BY2014 corpus.

The new recorded BY2014 corpus was about three times larger than the previous PB2007 corpus, and included mostly full sentences while the PB2007 focused on isolated words. Moreover, the number of EMA sensors used to record the articulatory data was increased. In particular, an additional sensor was placed on the soft palate. As expected, including this sensor resulted in a better synthesis accuracy for the nasal phones, in particular for the vowel /ɑ̃/ for which recognition accuracy jumped from 20% in the PB2007 synthesis (**Fig. 38**-A), to 58% in the BY2014 synthesis when not using glottal activity information (**Fig. 44**, first row), and 83% when using it (**Fig. 44**, second row), and for the consonant /m/, which accuracy jumped from 58% in the PB2007 synthesis (**Fig. 38**-B), to 81% and 92% in the BY2014 synthesis, when

respectively using and not using glottal activity information (**Fig. 44**, first and second rows). Using this additional glottal activity information resulted in the ability to discriminate pairs of consonants that mainly differed by their voicing features (such as /s/ and /z/). While the evaluation on vowel-consonant-vowel sequences suggested that synthesis must still be improved on plosive consonants, the synthesized speech was fully intelligible, as assessed by the transcription of synthesized sentences by human subjects (**Fig. 45**). As for the PB2007 corpus, dimensionality reduction techniques were applied to the articulatory data in order to observe how the synthesis intelligibility degrades when reducing the number of articulatory parameters. Results showed that using 14 articulatory parameters yields intelligibility scores that are close to the best scores achieved with 27 parameters, and that reducing further to 10 parameters did not significantly impact the intelligibility of consonants but slightly those of vowels. Even better results are to be expected if using deep auto-encoders instead of principal component analysis for dimensionality reduction. These results support the fact that synthesis of intelligible synthesis can be achieved with about a dozen of articulatory parameters.

Chapter 5: Real-time control of an articulatory-based speech synthesizer for silent speech conversion

I. Introduction

In the previous chapter we described an articulatory-based speech synthesizer using a deep neural network which maps articulatory data to acoustic features which are then converted to audible speech. Such synthesizer was built for an articulatory-acoustic data corpus, the BY2014 corpus, which was recorded from a specific speaker (in the following, the reference speaker). We showed that such speech synthesizer was able to produce fully intelligible synthesis from about a dozen of articulatory parameters, while being robust to noisy articulatory inputs. Moreover, the proposed synthesizer was compatible with real-time applications, such as a brain computer interface for speech rehabilitation.

However, it remains unknown whether a given articulatory-based speech synthesizer built from articulatory-acoustic data obtained in one particular reference speaker can be controlled in real time by any other speaker to produce intelligible speech. In this chapter, we thus assess how well the proposed synthesis approach can be used to produce intelligible speech when controlled by a speaker different than the reference speaker. This question is of particular importance in a BCI context in which the synthesizer will be controlled by another speaker, and using another modality (i.e. brain signals that will be decoded into articulatory parameters). Here we consider a simpler situation in which silent speakers (i.e. speakers that articulate but do not produce any sounds) are controlling this synthesizer using only the movement of their main speech articulators, in a closed-loop paradigm (**Fig. 46**). Such a silent speech condition is as close as possible to a speech BCI paradigm where the synthetic voice replaces the actual subject's voice. In this paradigm, a new speaker is equipped with EMA sensors, as when recording an articulatory-acoustic corpus. Using these EMA sensors, the articulatory movements of this subject are recorded while speaking silently, i.e. while articulating without actually producing any sound. These articulatory movements are then used to control the articulatory-based speech synthesizer presented in the previous chapter in order to produce the corresponding speech, which is played back to the silent speaker through earphones. Since the articulatory movements of this new speaker are different than those of the reference speaker, a calibration is needed in order to perform the articulatory-to-articulatory mapping, i.e. in order to convert the articulatory movements of the new speaker into articulatory parameters suitable for the synthesizer.

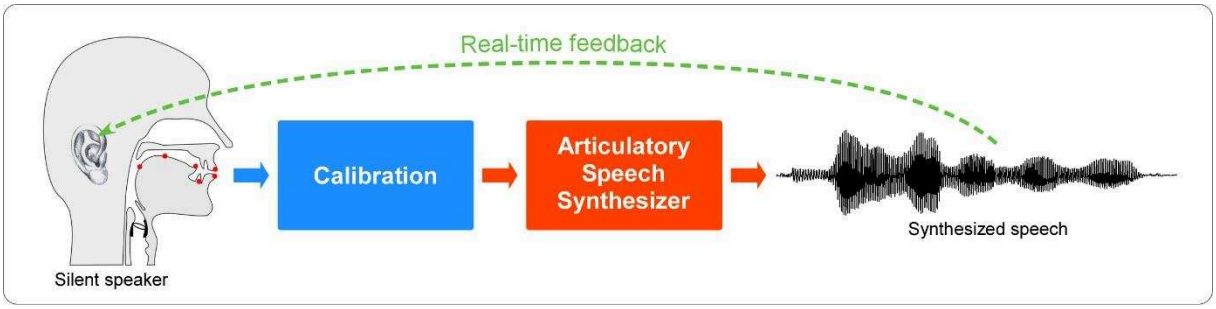


Fig. 46: Real-time closed loop paradigm. Articulatory data from a silent speaker are recorded and converted into articulatory input parameters for the articulatory-based speech synthesizer. The speaker receives the auditory feedback of the produced speech through earphones.

Several studies have addressed the problem of "silent speech recognition", i.e. identifying a word sequence from a silent articulation, under different modalities including ultrasound (Hueber et al., 2009), electromyography (Wand and Schultz, 2011), non-audible murmur (Heracleous et al., 2004), or permanent-magnetic articulography (Gilbert et al., 2010). Other studies have addressed the problem of "silent speech conversion", i.e. directly reconstructing a synthetic speech signal from silent articulation, without any restriction on the vocabulary, for instance with ultrasound (Hueber and Bailly, 2016) or with electromyography (Janke et al., 2012; Wand et al., 2013). Here we consider silent speech conversion, in which we process in real-time an EMA data flow in order to synthesize speech (Bocquelet et al., 2015, 2016a).

II. Methods

1. Subjects and experimental design of the real-time closed-loop synthesis

Four subjects (1 female, 3 males) controlled the synthesizer in real time. The reference speaker was one of them (Speaker 1), i.e. he was also used as a test subject, but with data from a different session than the reference data session. The whole experimental protocol is summarized in **Fig. 47**.

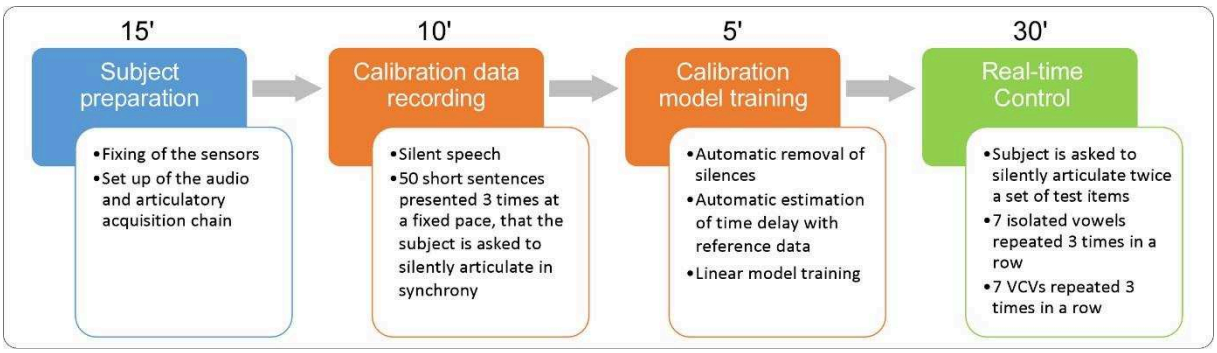


Fig. 47: Experimental protocol for the real-time closed-loop synthesis. First, sensors are glued on *the speaker's articulators,* then articulatory data for the calibration is recorded in order to compute the articulatory-to-articulatory mapping, and finally the speaker articulates a set of test items during the closed-loop real-time control of the synthesizer.

First, EMA sensors were glued to the subject main articulator, using the same approach as for recording the BY2014 articulatory-acoustic corpus. The articulatory data was recorded using the NDI Wave system in similar conditions as for acquiring the reference data (recording

at 400Hz and down-sampling to 100Hz), except that only 6 sensors were used to record the articulatory movements of the lower and upper lips, the tongue tip, dorsum and back, and the jaw. This was due to the fact that the NDI Wave system was limited to 6 sensors when retrieving the articulatory data in real-time (while it was possible to use more sensors in offline mode). The soft palate sensor was one of the discarded sensor because most subjects (3 out of 4) were very uncomfortable with keeping the soft palate sensor for a long duration. While the coils were positioned at the same anatomical locations, no particular attention was given to place them at very precise locations.

In a second step, calibration data was recorded in order to compute the so-called "articulatory-to-articulatory" mapping model, which allows to map articulatory data from a new speaker to the articulatory space of the reference speaker. In order to estimate the articulatory-to-articulatory mapping, it was necessary to obtain articulatory data from the new speakers in synchrony with articulatory data from the reference speaker when articulating the same sounds. The new speakers were thus asked to silently repeat a subset of 50 short sentences (about 4 words each), extracted from the reference BY2014 corpus, in synchrony with the corresponding audio presented through earphones. Each sentence was first displayed on the screen during one second, then after a visual countdown, it was played three times at a fixed pace, so that the speaker could adapt to the reference speaker rate and way of speaking. Only the last repetition was considered in order to obtain the best temporal synchronization. Indeed, preliminary experiments (not reported here) showed that synchronization was crucial in order to achieve a good articulatory-to-articulatory mapping. Subjects were asked to repeat the sentences silently, and not loudly, because of significant differences that exist between silent and vocalized speech (Hueber et al., 2010a; Janke et al., 2010), and because subsequent real-time closed-loop control of the synthesizer would then be achieved while subjects were silently speaking. Here we made the choice for a supervised calibration approach, i.e. a calibration that needs to record a specific corpus of data, but that other approaches could be considered, including unsupervised approach which do not need to record additional specific data (Wand and Schultz, 2014). Our choice was motivated in order to be as close as possible to BCI calibration methods, which are mostly supervised. This explains also why we did not use a more advanced method to align the new speaker's audio on the reference speaker's audio, such as dynamic time warping, but rather asked the new speaker to align his speech production on the reference audio. Indeed, in the case of BCI, the patient will not be able to produce speech and thus no audio would be available to perform a better alignment.

In a third step, the calibration data was used to train the articulatory-to-articulatory mapping model, after automatically aligning the new articulatory data to the reference data, and automatically removing silences since EMA data does not contain any information about vocalization but only about the articulatory configuration for the subject.

Finally, this calibration model was then applied in real time to incoming articulatory trajectories of each silent speaker to produce continuous input to the speech synthesizer, in this case the BY2014 speech synthesizer. Since the subjects were in a silent speech condition and thus no glottal activity was available, we chose to perform the synthesis using the fixed-pitch template-based excitation, and in order to reduce the number of control parameters, we chose the synthesis model based on 14 articulatory parameters which showed that it was able to produce fully intelligible speech (see "Conclusion" in the previous chapter). During this real-

time control of the synthesizer, the subjects were asked to pronounce a set of test items in order to assess the synthesis intelligibility.

2. Articulatory-to-articulatory mapping

As built, the synthesizer could only be used on the reference data and could not be directly controlled by another speaker or even by the same speaker in a different session. Indeed, from one session to another, sensors might not be placed at the exact same positions with the exact same orientation, or the number of sensors could change, or the speaker could be a new subject with a different vocal tract geometry and different ways of articulating the same sounds. In order to take into account the differences between the reference speaker and a new speaker, it was necessary to calibrate a mapping from the articulatory space of each new speaker (or the same reference speaker in a new session) to the articulatory space of the reference speaker, that is, an articulatory-to-articulatory mapping (**Fig. 48**, left blue part).

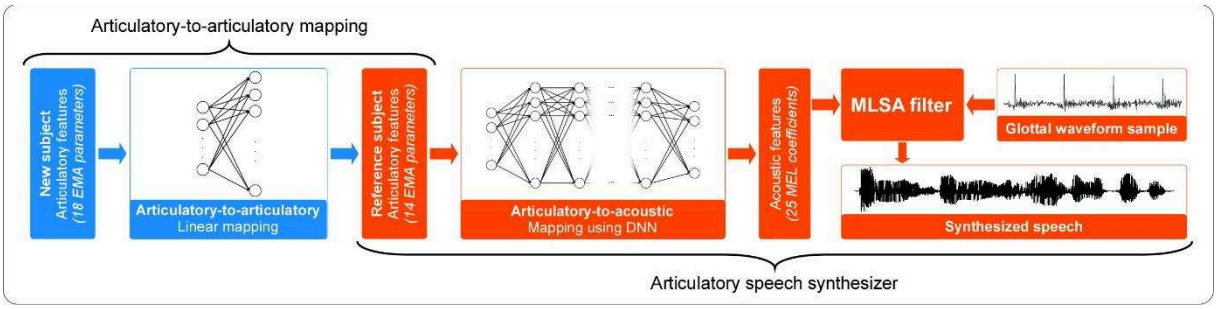


Fig. 48: Processing chain for real-time closed-loop articulatory synthesis. The articulatory-to-articulatory (left part) and articulatory-to-acoustic mappings (right part) are cascaded. Items that depend on the reference speaker are in orange, while those that depend on the new speaker are in blue. The articulatory features of the new speaker are linearly mapped to articulatory features of the reference speaker, which are then mapped to acoustic features using a DNN, which in turn are eventually converted into an audible signal using the MLSA filter and the template-based excitation signal.

For each silent speaker, the articulatory-to-articulatory mapping was performed using a linear model mapping the articulatory data of the speaker to those of the reference speaker. The choice for a linear model was motivated by preliminary experiments – not reported here – that showed no advantage of using more complex models such as artificial neural networks, probably because of the limited amount of calibration data. Note the velum trajectory taken as input to the synthesizer was predicted from tongue, lips and jaw trajectories as no sensor was placed on the soft palate for the four subjects.

In order to counterbalance speaker and system latencies, a global delay between new and reference articulatory data was estimated for each speaker by trying different delays. A different linear model between the new speaker's articulatory data and the reference data was computed for each candidate delay. Each linear model was then applied to the new speaker's articulatory data, and the mean-squared error (MSE) between predicted and actual reference articulatory data was computed. The delay which led to the smallest MSE was considered to be due to system latency and was corrected before the training of the final model by simply shifting and cutting the original data. Frames corresponding to silence periods were then discarded using the aligned transcription from the BY2014 corpus.

Out of the 50 calibration sentences, 40 were used for the training of the articulatory-to-articulatory mapping while the remaining 10 sentences were kept for evaluation (randomly

chosen, but identical across all speakers). By contrast with the articulatory-to-acoustic mapping, the articulatory-to-articulatory mapping was done frame-by-frame, i.e. without concatenating any past frame.

3. Implementation details

As shown in **Fig. 48**-B, the real-time control of the articulatory-based speech synthesizer was achieved by cascading the linear model used for articulatory-to-articulatory mapping and the DNN used for articulatory-to-acoustic mapping. Here we used the DNN trained on the BY2014 corpus, with 3 hidden layers of 200 units each, which maps 14 articulatory parameters to 25 mel coefficients. The input articulatory data capture and processing (especially the re-referencing with regards to the reference sensor), the linear and DNN mappings and the MLSA filter were all implemented within the Max/MSP environment (Cycling'74, Walnut CA, USA, https://cycling74.com/products/max/) dedicated to real-time audio processing. The integration of the DNN and linear mappings was done through the dedicated C++ library that we developped (see previous chapter).

Special attention was given to audio settings in order to minimize the audio chain latency and obtain a delay inferior to 30ms. This way, the silent speaker could rely on the synthetic speech as an auditory feedback and exploit it to regulate his own production. According to the literature on delayed auditory feedback (Lincoln et al., 2006), the latency should be no greater than about 50 ms. A larger latency might generate a conflict between kinesthesic and auditory feedbacks.

Since the subjects were in silent speech condition, and thus no glottal activity was present, we used the template-based excitation signal for the MLSA filter.

4. Closed-loop experimental paradigm

During this closed-loop situation, each speaker was asked to silently articulate a set of test items while given the synthesized auditory feedback. This auditory feedback was recorded for further intelligibility evaluation. Subjects were allowed to adapt to the closed-loop situation for at least 20 minutes. During this closed-loop situation, each speaker was silently articulating and given the synthesized auditory feedback through amagnetic Nicolet TIP-300 insert earphones (Nicolet Biomedical, Madison, USA) ensuring no interference with the magnetic field of the NDI Wave system. Then, they were asked to pronounce a set of test items, which were not part of the datasets used to train the articulatory-to-acoustic and the articulatory-to-articulatory mappings.

The test set consisted of the 7 isolated vowels /a/, /e/, /i/, /o/, /u/, /œ/, and /y/, and 21 vowel-consonant-vowel (VCV) pseudo-words made by the 7 consonants /b/, /d/, /g/, /l/, /v/, /z/ and /ʒ/, in /a/, /i/ and /u/ context (e.g., 'aba' or 'ili'). We chose not to include nasal vowels (e.g., /ɑ̃/, which corresponds to a nasalized /a/) since no sensor was placed on the soft palate. Likewise, we did not include nasal consonants (e.g., /m/ or /n/) and most unvoiced consonants (e.g., /p/ which roughly corresponds to an unvoiced /b/). Each item of the test set was repeated three times in a row. The whole set of items was repeated three times by each speaker, each repetition

being separated by about 10 minutes of free control of the synthesizer. Each isolated vowel or VCV was thus repeated 9 times by each subjects.

5. Evaluation of the synthesis quality

The quality of the synthesized sounds was assessed in two ways. We first carried out a qualitative evaluation, in which the acoustic signals were compared with the original signals processed through analysis-synthesis and with the offline synthesis using the reference data (referred here as the "Reference offline synthesis"). Analysis-synthesis was performed by converting the audio signals into mel-cepstrum (MEL) coefficients, which were then directly converted back into audio signals using the MLSA filter and template-based excitation. Such signal is referred here to as the "anasynth signal". This conversion is not lossless, though it represents what would be the best achievable quality for the synthetic speech signal in the present context.

Then we carried out a quantitative evaluation of our system through an intelligibility test, similarly to the way the reference offline synthesis was evaluated (see "Subjective evaluation using listening tests" in previous chapter). Twelve subjects participated to this test. All participants were French native speakers with no hearing impairment. Each listener evaluated 3 repetitions (randomly picked for each listener) of each of the 28 test items for each of the 4 new speakers. Remind that for the real-time closed-loop synthesis, the stimuli were generated using only the fixed-pitch template-based excitation. In total, each listener had thus to identify 336 sounds (7 vowels + 21 VCVs, three times for each of the 4 speakers). The sounds were all normalized using automatic gain control, and played in random order at the same sound level through Beyerdynamic DT-770 Pro 80 Ohms headphones, while the listener was seated in a quiet environment. No performance feedback was provided during the test. Participants were instructed to select from a list what they thought was the corresponding vowel in the case of an isolated vowel, or the middle consonant in the case of a VCV sequence. Graphical user interface buttons were randomly shuffled for each subject in order to avoid systematic default choice (e.g., always choosing the left button when unable to identify a sound). The subjects were told that some of the sounds were difficult to identify, and thus to choose the closest sound among the offered possibilities. The recognition accuracy was defined as $Acc\% = 100\frac{R}{N}$with R the number of correct answers for the N presented sounds of the test. Thus, the chance level was $1/7 \approx 14\%$ both for vowels and VCVs.

6. Statistical analysis

a. Analysis of the articulatory-to-articulatory mapping

For the articulatory-to-articulatory mapping, mean distance between predicted articulatory trajectories and reference articulatory trajectories was computed for each item of the test corpus and each speaker. A two-factor ANOVA with repeated measures was performed using the following model: Distance ~ Sensor*RefSpeaker + Error(Item / (Sensor*RefSpeaker)), where Distance is the mean distance between predicted and reference trajectories, RefSpeaker has two levels indicating if it was the reference speaker (Speaker 1) or another speaker (Speaker 2, 3 or

4), Item corresponds to the identifier of the tested item, and Sensor has 7 levels corresponding to the different EMA sensor positions to be predicted (upper lip, lower lip, jaw, tongue tip, tongue dorsum, tongue back and velum). Multiple comparisons were made using contrasts according to (Hothorn et al., 2008). All the tests were made using the R software, and packages lme4, and multcomp.

b. Analysis of the real-time closed-loop synthesis

As for offline reference synthesis (see "Statistical analysis" in previous chapter), the statistical analysis of the real-time closed-loop synthesis results was performed using mixed logistic regression. The following model was used: Result ~ (Segment + RefSpeaker) ^ 2 + (1 | Listener), where Result is the binary answer (equals 0 if the item was wrongly identified, otherwise 1), Segment has two levels corresponding to the type of item (vowel or VCV), RefSpeaker has two levels indicating if it was the reference speaker (Speaker 1) or another speaker (Speaker 2, 3 or 4), and Listener has 12 levels corresponding to each listener that participated in the listening test. Multiple comparisons were made using contrasts according to (Hothorn et al., 2008). All the tests were made using the R software, and packages lme4, multcomp and lsmeans.

III. Results

1. Accuracy of the articulatory-to-articulatory mapping

Fig. 49-A shows an example of articulatory data recorded from a new speaker (from Speaker 2), with the corresponding reference audio signal that the speaker was presented and asked to silently repeat synchronously (in this example, the sentence was *"Deux jolis boubous"*, meaning *"two* nice *booboos"*, which was not part of the training set). **Fig. 49**-B shows the transformation of these signals after their articulatory-to-articulatory mapping onto the reference speaker's articulatory space. One can clearly see that articulatory movements of the new speaker were originally quite different than those of the reference speaker; and that they became similar once the articulatory-to-articulatory mapping was performed.

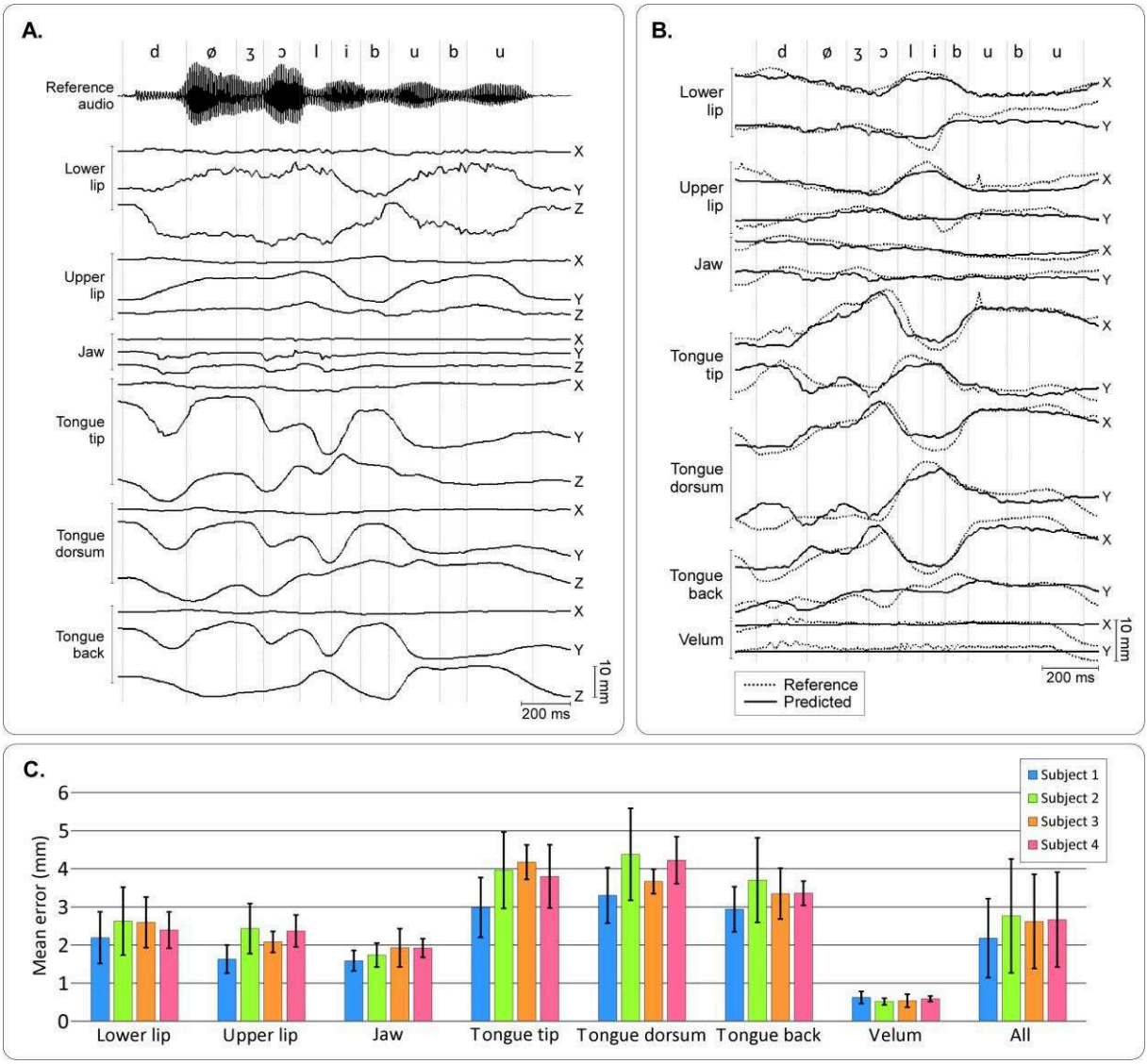


Fig. 49: Articulatory-to-articulatory mapping. A – Articulatory data recorded from a new speaker (Speaker 2) and corresponding r*eference audio signal for the sentence "Deux jolis boubous" ("Two* nice *booboos"). For each sensor, the X* (rostro-caudal), Y (ventro-dorsal) and Z (left-right) coordinates are plotted. Dashed lines show the phonetic segmentation of the reference audio, which the new speaker was ask to silently repeat in synchrony. B – Reference articulatory data (dashed line), and articulatory data of Speaker 2 after articulatory-to-articulatory linear mapping (predicted, plain line) for the same sentence as in A. Note that X, Y, Z data were mapped onto X, Y positions on the midsagittal plane. C – Mean Euclidean distance between reference and predicted sensor position in the reference midsagittal plane for each speaker and each sensor, averaged over the duration of all speech *sounds of the calibration corpus. Error bars show the standard deviations, and "All" refer to* mean distance error when pooling all the sensors together.

We further quantified the quality of the articulatory-to-articulatory mapping. Since articulatory data consists of geometrical coordinates, the mean Euclidean distance between predicted and true positions could be estimated for each sensor and for each speaker (**Fig. 49**-C). The average error across all sensors and speakers was 2.5 mm ± 1.5 mm. Errors were significantly higher for tongue sensors than for non-tongue sensors ($P < 0.005$ for 22 out of 24 pairwise comparisons corrected for multiple comparisons – see "Analysis of the articulatory-to-articulatory mapping"), and lower for the velum sensor than for the non-velum sensors ($P < 0.001$ for 10 out of 12 pairwise comparisons corrected for multiple comparisons – see "Analysis of the articulatory-to-articulatory mapping"). This is consistent with the fact that the tongue and velum are the articulators for which movement amplitudes were the highest and lowest,

respectively (see **Fig. 19**-B). Mean distances for the reference speaker (Speaker 1) were systematically lower than for other speakers for all sensors except the velum. These differences were statistically significant for the tongue tip (P = 0.00229) and the tongue dorsum (P = 0.03051).

2. Intelligibility of the real-time closed-loop synthesis

During real-time control, the speakers were asked to reproduce a specific set of test sounds. The remaining time of the experiment was kept for other tasks, including spontaneous conversations. **Fig. 50** shows examples of spectrograms of vowels and VCVs obtained during a session of real-time control (Speaker 2, first occurrence of each sound), compared with the corresponding spectrograms of anasynth and reference offline synthesis sounds. In general, we found that the spectrograms for the three conditions presented very similar characteristics, although some differences did exist in their fine structure, especially for consonants. For instance, the real-time examples of the plosive consonants /b/, /d/ and /g/ showed more energy smearing from vocalic to consonant segments as compared to the anasynth and offline synthesized versions. Also, the real-time example of /ʒ/ had characteristics closer to the anasynth version of /l/ than to the anasynth version of /ʒ/ (**Fig. 50**-B).

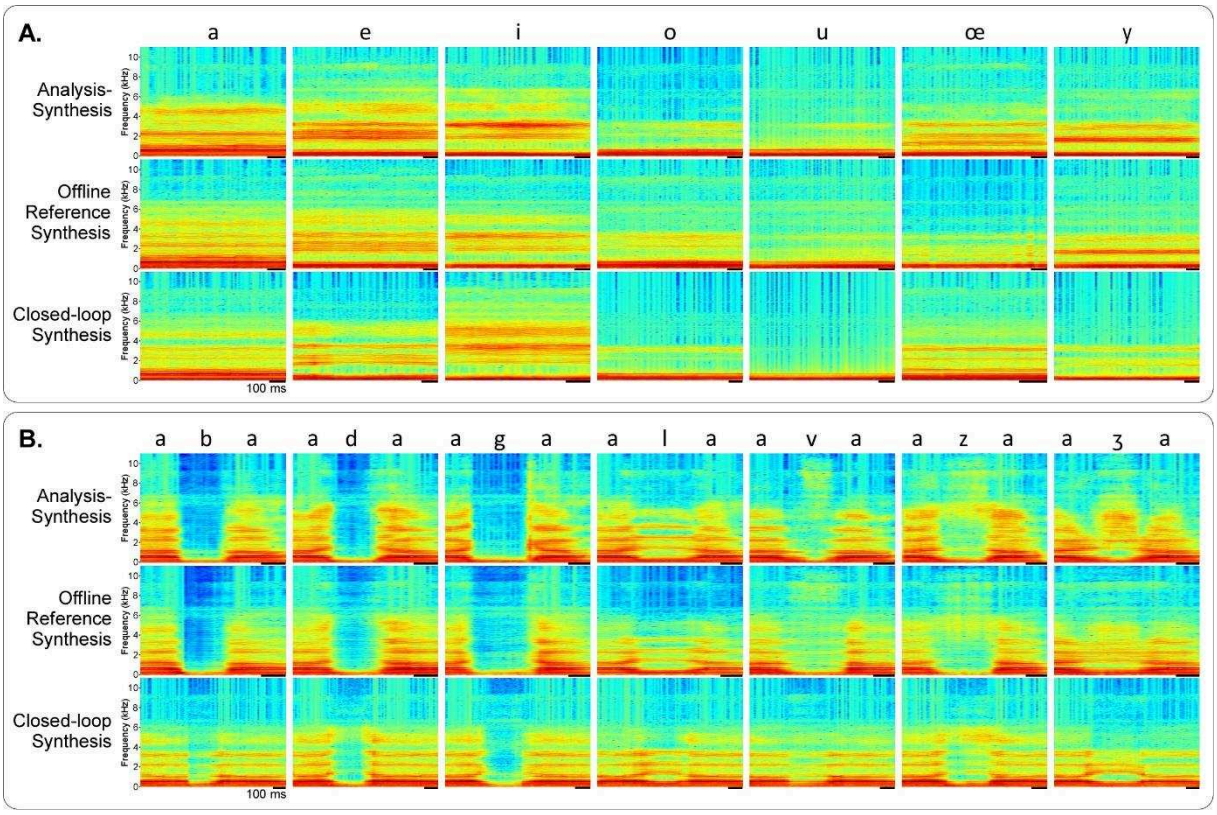


Fig. 50: Real-time closed loop synthesis examples. Examples of audio spectrograms for anasynth, reference offline synthesis and real-time closed-*loop (Speaker 2), for the vowels /a/, /e/, /i/, /o/, /u/, /œ/ and /y/ (A), and for the conso*nants /b/, /d/, /g/, /l/, /v/, /z/ and /ʒ/ in /a/ context (B). The thick black line under the spectrograms corresponds to 100 ms.

The test sounds produced in the closed-loop experiment were recorded and then their intelligibility was evaluated in the same way as for the offline synthesis intelligibility evaluation, i.e. a subjective intelligibility test performed by 12 listeners. **Fig. 51** summarizes the results of this listening test. The speech sounds produced by all 4 speakers obtained high

vowel accuracy (93% for Speaker 1, 76% for Speaker 2, 85% for Speaker 3, and 88% for Speaker 4, leading to a mean accuracy score of 86%), and reasonable consonant accuracy (52% for Speaker 1, 49% for Speaker 2, 48% for Speaker 3, and 48% for Speaker 4, leading to a mean accuracy score of 49%). These scores were far above chance level (chance = 14%, $P < 0.001$) for both vowels and consonants. For all speakers, the 48-52% VCVs accuracy obtained during real-time control is to be compared to the 61% score obtained for the same VCVs in the offline reference synthesis. The difference is statistically significant ($P = 0.020$ for reference speaker and $P < 0.001$ for other speakers, compare **Fig. 51**-A and **Fig. 43**-A) but the decrease is quite limited when considering that the speaker is no longer the reference speaker and that the synthesis is performed in an online closed-loop condition. The same observation applies to the vowel identification results: the 76-93% vowel accuracy for the closed-loop online synthesis is also found significantly lower than the 99% accuracy score obtained for the same vowels in the offline synthesis ($P < 0.001$ for reference and other speakers), but the decrease is relatively limited. The recognition accuracy for vowels was significantly higher for the reference speaker ($P = 0.002$) but no significant difference between the reference speaker and the other speakers was found for the VCVs ($P = 0.262$), even if the reference speaker obtained the highest average accuracy value for VCVs.

Regarding the vocalic context (**Fig. 51**-B), VCVs in /a/ context had better recognition accuracy than those in /i/ ($P < 0.001$) and /u/ ($P < 0.001$) contexts for all subjects, which is consistent with results from the offline reference synthesis (**Fig. 43**-B). VCVs in /u/ context were found to have a better recognition accuracy than those in /i/ context ($P = 0.009$). Regarding the VCVs (**Fig. 51**-C), the recognition accuracy varied largely across consonants, ranging from an average of 21% for /b/ to 85% for /ʒ/. It was generally lower for the plosive consonants /b/, /d/ and /g/, which is consistent with results from the offline reference synthesis, while the accuracy on the remaining consonants was different for each subject. For instance, Subjects 1, 2 and 3 had good accuracy on /v/ while Subject 4 had a much lower accuracy. Similar result can be observed for /z/ and /ʒ/ for different subjects.

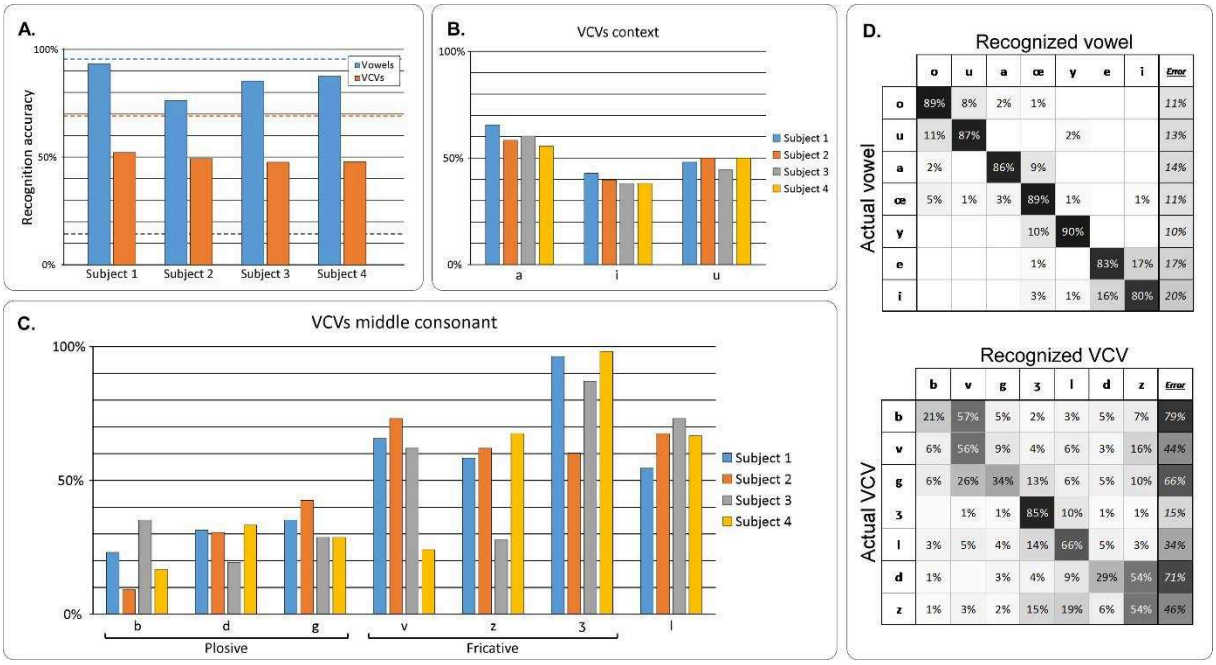


Fig. 51: Results of the subjective listening test for real-time articulatory synthesis. A – Recognition accuracy for vowels and consonants, for each subject. The grey dashed line shows the chance level, while the blue and orange dashed lines show the

corresponding recognition accuracy for the offline articulatory synthesis, for vowels and consonants respectively (on the same subsets of phones). B – Recognition accuracy for the VCVs regarding the vowel context, for each subject. C – Recognition accuracy for the VCVs, by consonant and for each subject. D – Confusion matrices for vowels (left) and consonants from VCVs in /a/ context (right). Rows correspond to ground truth while columns correspond to user answer. The last column indicates the amount of errors made on each phone. Cells are colored by their values, while text color is for readability only.

Confusion matrices for both vowels and consonants are shown in **Fig. 51**-D. These confusion matrices present features that are similar to the confusion matrices obtained for offline articulatory synthesis (**Fig. 44**), and summarize the results quality. All vowels show a recognition accuracy above 80%, and the highest accuracy was obtained for /y/, with 90%. The majority of the confusions are between /e/ and /i/ (17% of /e/ were recognized as /i/, and 16% of /i/ as /e/). Secondary confusions are between /o/ and /u/ (11% of /u/ were recognized as /o/, and 8% of /o/ as /u/), between /y/ and /œ/ (10% of /y/ were recognized as /œ/), and between /a/ and /œ/ (9% of /a/ were recognized as /œ/). The confusion matrix for consonant roughly corresponds to the confusion matrix obtained for offline articulatory synthesis, with emphasized confusions. Thus, the main confusions occurred again for plosive consonants /b/ (57% of /b/ were recognized as /v/) and /d/ (54% of /d/ were recognized as /z/), while quite few errors were made on /ʒ/ (85% of accuracy). Some errors were also made on /g/ but with less systematic confusion (26% with /v/, 13% with /ʒ/, and 10% with /z/). However, new confusions appeared that explain the significant drop in consonants accuracy with respect to offline articulatory synthesis: between /ʒ/ and /l/ (10% of /ʒ/ were recognized as /l/), and between /z/ and /l/ (19% of /z/ were recognized as /l/).

As previously mentionned, subjects had several minutes of free control of the synthesizer between two consecutive repetitions of the test items. This was purposely chosen in order to observe if subjects could fastly learn and improve their control of the speech synthesizer. **Fig. 52** shows the evolution of the recognition accuracy before and after training, i.e. from the first repetition of the test items to the last one (approximately separated by 20 minutes).

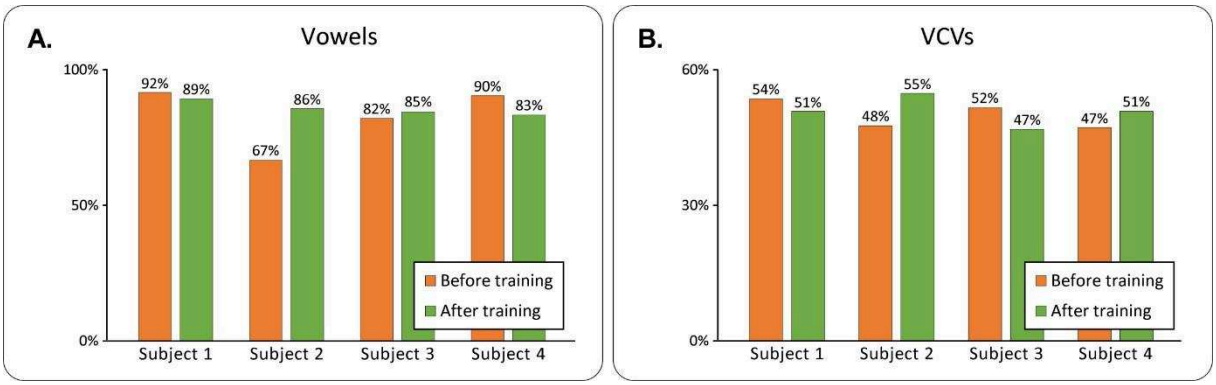


Fig. 52: Evaluation of the real-time closed-loop synthesis before and after subjects training. A – Recognition accuracy for vowels, before and after a short training time, for each subject. B – Recognition accuracy for VCVs, before and after a short training time, for each subject.

Interestingly, there was no clear effect of training time: while Subject2 reported to improve during training, which was confirmed by the data both on vowels and VCVs, no clear improvement could be observed on the results from other subjects. For instance, the recognition accuracy slightly increased for vowels after training, but slightly decreased for VCVs, while the opposite can be observed for Subject 4.

3. Spontaneous conversations

During the free control of the speech synthesizer, episodes of spontaneous conversations between the subject and the experimenter could be achieved, in particular with Subject 1 (which is the reference speaker), and Subject 2. Note that the experimenter could not see the subjects, which were in a separate room, but could only ear the synthesized feedback through headphones. However, the experimenter was familiar with the synthesis and thus better at discriminating the synthesized sounds than a naïve subject. Spontaneous conversations ranged from simple "yes" / "no" answers, to complete sentences such as "I am gonna be a father". Some of these sentences were manually transcribed during the experiment and are reported in Annex 2. However, they were not recorded during the experiment so that they were not evaluated and are thus only provided for illustratory purpose. Future experiments should include full sentences in the test items in order to properly evaluate the synthesis intelligibility on sentences during the real-time control.

IV. Conclusion on the real-time control of the articulatory-based speech synthesizer

In this chapter we showed that the articulatory-based speech synthesizer presented in the previous chapter could be controlled in real-time closed-loop situation by several speakers using motion capture data (electromagnetic articulography) as input parameters. Experiments included the same reference speaker in a different session, as well as other speakers. All speakers were silently articulating and were given the synthesized acoustic feedback through headphones. A calibration method was used to take into account articulatory differences across speakers (and across sessions for the reference speaker), such as sensor positioning and ways of articulating the different sounds. Subjective listening tests were conducted to assess the synthesis intelligibility during real-time closed-loop control by new speakers.

The phone recognition accuracy was far above chance level, both for vowels and consonants (**Fig. 51**). Interestingly, this good intelligibility was obtained despite significant trajectory errors made on input control parameters obtained by the articulatory-to-articulatory mapping (about 2.5 mm on average, see **Fig. 49**-B). This confirms the previous results indicating that DNN-based articulatory synthesis is robust to fluctuations of the input parameters (see previous chapter). As expected, the closed-loop synthesis intelligibility was lower than for the reference offline synthesis. However, it was relatively limited. Confusions were similarly distributed in both cases, indicating that using the synthesizer in a closed-loop paradigm mainly emphasized the already existing confusions. The fact that most errors were consistent between offline and closed-loop synthesis suggests that real-time closed-loop articulatory synthesis could still benefit from improving the articulatory-to-acoustic mapping. This could be achieved by efficiently detecting specific constrictions from the articulatory data in order to improve the synthesis of plosive consonants, which are the major source of errors. The presence of additional minor confusions suggests that other aspects might also be improved, such as the articulatory-to-articulatory mapping with a better calibration approach. Indeed, to remain in a situation as close as possible to future BCI paradigms with aphasic participants, the articulatory-to-articulatory calibration step was performed under a silent

speech condition. This was also consistent with the fact that the closed-loop condition was also performed in a silent speech condition so that the speaker received only the synthesized feedback, not superimposed on its own produced speech. Thus the articulatory-to-articulatory mapping converted articulatory trajectories recorded under a silent speech condition (for each speaker) into articulatory trajectories recorded under overt speech condition (of the reference speaker). Previous studies have shown that articulatory movements differ between silent and overt speech, and especially that silent speakers tend to hypo-articulate (Hueber et al., 2010a; Janke et al., 2010). Such phenomenon may thus leads to smaller discrimination of articulatory trajectories during silent speech.

Improving the articulatory-to-articulatory and the articulatory-to-acoustic mappings might however not be the sole possibility to improve the intelligibility of closed-loop speech synthesis. Indeed, while results from the evaluation of the articulatory-to-articulatory mapping showed that for most sensors the mean prediction error was lower for Speaker 1 (the reference speaker), the results obtained during the real-time experiment showed that other speakers could achieve a control of the articulatory synthesizer similar to Speaker 1, in particular for consonants (see **Fig. 51**-A). For example, episodes of spontaneous conversation could be achieved not only with Speaker 1 but also with Speaker 2 (see "Spontaneous conversations"). This suggests that other factors come into play for the control of the synthesizer. One possibility is that subjects may adapt differently to the articulatory-to-articulatory mapping errors and find behavioral strategies to compensate for these errors. Here, each subject had about 20 minutes of free closed-loop control of the synthesizer between the two productions of test items, but we could not see any significant improvement over this short period of time (see "Intelligibility of the real-time closed-loop synthesis"). Finding behavioral strategies might thus need a more significant amount of training time.

Altogether, these results show that an intelligible articulatory-based speech synthesizer can be controlled in real-time by different speakers to produce not only vowels, but also intelligible consonants and some sentences. This synthesizer built from a reference speaker data in an overt speech condition could be controlled to produce free speech in real time in a silent speech condition by other speakers with a different vocal tract anatomy and a different articulatory strategy using a simple linear calibration stage. Note that while this is not the topic of this manuscript, this result is of particular interest for the emerging research field on "silent speech interfaces", which are lightweight devices able to capture silent articulation using non-invasive sensors and convert it into audible speech (Denby et al., 2010; Wand et al., 2013; Cler et al., 2014; Hueber and Bailly, 2016). Indeed, although the presented EMA-based interface is not strictly a silent-speech interface, the present results indicate that it is possible to synthesize intelligible speech in real time from articulatory data acquired in silent speech condition. Further studies could extend these results using less invasive techniques to obtain articulatory signals, such as EMG (Wand et al., 2013; Cler et al., 2014) and/or ultrasound signals (Hueber and Bailly, 2016).

Discussion on the articulatory-based speech synthesis for BCI applications

In **Chapter 3**, we first presented a new articulatory-acoustic dataset, the BY2014 corpus, recorded from a single French male subject using electromagnetography (EMA). While some EMA datasets were already pubicly available in other languages, such as the MOCHA dataset in english, the BY2014 is the first French EMA dataset pubicly available.

We then showed in **Chapter 4** that this dataset could be used to create an articulatory-based speech synthesizer, using deep neural networks, that could produce fully intelligible speech. Indeed, while the best synthesis got about 70% accuracy on vowels and consonants, the word recognition accuracy on full sentences was still above 90%. This is a well-known phenomenon in the automatic speech recognition field where recognition models generally have a non-perfect recognition of individual phones, but exploit a priori information, such as a limited vocabulary or a defined language, to improve the recognition accuracy at word and sentence level. Similar approach could be envisioned to enhance the synthesis intelligibility by exploiting such a priori information. This could be done by combining the advantage of deep neural networks for regression of continuous variables, with the advantages of hidden markov models for modeling sequences of discretes states, such as phones, words or semantic units. Moreover, results showed that information about the glottal activity is crucial in order to discriminate pairs of consonants that mainly differ by their voicing feature. Indeed, removing the glottal activity information – in our case the pitch – resulted in a drastic decrease of recognition accuracy for the six unvoiced consonants, becoming close to null. In a BCI paradigm, we could envision to use the brain activity to predict the pitch, or at least predict if sounds are voiced or unvoiced in a binary fashion, which should be enough to discriminate unvoiced consonants from their voiced counterparts. Several studies reported laryngeal specific activity in the speech motor cortex (Brown et al., 2008; Grabski et al., 2012; Bouchard et al., 2013), and one BCI study suggested that the voicing feature could be decoded from brain activity (Lotte et al., 2015). However, predicting the voicing from the brain activity remains challenging considering that phones have short duration, of about 10 to 100ms, so that the prediction of the voicing feature has to be as fast and precise, probably needing a time resolution below 10ms. We as well showed that this synthesizer was robust to noisy articulatory inputs, and could rely on few articulatory parameters to produce speech, which are both potential assets for BCI that is known to produce unperfect control parameters, with a limited number of degrees of freedom. In particular, results showed that reducing articulatory parameters from 27 to 10 parameters did not significantly impact the intelligibility of consonant, but rather that of vowels. This result is quite unexpected given that consonants consist in more complex articulatory movements, for which timing is crucial, while vowels consist more of static articulatory positions. This should be further investigated in future studies. A possible solution would be to relate the reduced articulatory components with the original articulatory features in order to identify articulators that were mainly preserved or discarded.

Finally, we showed in **Chapter 5** that this synthesizer that was build from a specific subject, could be controlled in real-time, in a closed-loop paradigm, by several subject, different than the original subject on which it was built. This further confirmed the results from **Chapter 4**, showing that the proposed DNN-based synthesis was robust to noisy articulatory inputs.

Indeed, real-time closed-loop control was possible while predicted articulatory control parameters had non-negligible errors of about several millimeters. Nonetheless, the real-time closed-loop synthesis quality was below that of the offline synthesis and on a limited set of phones, even after a training period of about half an hour. Such a short training time might not be sufficient for the subjects to find behavioral strategies in order to improve the synthesis quality, and several days might be needed to observe such improvement. Further studying the possibilities of improvement through training would be particularly interesting for BCI applications, for which it has been shown that consequent training time – about several days or weeks – is crucial in order to achieve accurate control by BCI. This would require to find a way to record stable EMA signals between sessions, since the EMA sensors can only be kept for about an hour or two. Unsupervised calibration methods, such as the one used in (Wand and Schultz, 2011) in the case of electromyographic recordings, could be considered to avoid this difficulty. In the case of a BCI, this would require to record stable neural signals over several days or weeks, which remains a challenging tasks when recording individual neurons. Finally, it is worth mentioning that all subjects had to control the same articulatory-based speech synthesizer, which was built from a single subject, with his own specific articulatory configuration.

Altogether, these results are a first step toward future speech BCI applications. Here we indeed showed that closed-loop speech synthesis was possible by subjects that had different speech production constrains (e.g., different anatomy of the vocal tract, articulatory idiosyncrasy) than those of the reference speaker from whom the speech synthesizer was built. This means that differences in anatomical constrains could be compensated by the articulatory-to-articulatory mapping. In the context of a speech BCI paradigm, a similar situation will be encountered, where the synthesizer will be built from a subject different from the BCI participants. In this case, the question will be whether differences in neuronal constrains between individuals can also be compensated by a proper neural signal decoding strategy. This is particularly true when recording from a small cortical area in which the neural activity is better correlated to some speech articulators than others. Micro-electrodes arrays that are used for recording individual neurons typically cover about 10mm² of the cortical surface, while the speech motor cortext area is about several cm². Thus, the positioning of the recording electrode array must be optimized in order to capture informative speech-related brain activity. Different positioning of the electrodes array could result in decoding some articulatory trajectories better than others. Here, the DNN-based mapping approach was robust to trajectory errors of several millimeters that were present in the input signals of the synthesizer resulting from imperfections in the articulatory-to-articulatory mapping. This is encouraging given that decoding neural signal into control parameters of the synthesizer will also be imperfect, and suggests that an articulatory-based speech synthesizer such as the one developed here is a good candidate for being used in a speech BCI paradigm. In this manuscript, we made the choice to envision articulatory parameters as an intermediate representation for decoding speech from neural activity recorded from the speech motor cortex. This hypothesis has to be tested in BCI experiment and compared to a direct decoding of cortical activity into acoustic speech parameters (Herff et al., n.d.). Preliminary experiments toward this goal are presented in the next part of this thesis.

Part 4: Thesis result 2 – Toward a BCI for speech rehabilitation

As previously mentioned, the goal of this thesis was to set the ground for a Brain Computer Interface (BCI) for speech restauration, in which neural activity is recorded from the speech motor cortex and decoded in control parameters for an articulatory-based synthesizer. In the previous part of this thesis, we thus presented an articulatory-based speech synthesizer which can be used to synthesize speech from articulatory data. Such articulatory data could be obtained by decoding the neural activity of a patient intending to speak. However, to date, there has not yet been any demonstration of an open-vocabulary BCI able to reconstruct continuous intelligible speech in real-time. One goal of this thesis was thus to make a first step toward such a speech BCI by investigating the decoding of speech and articulatory features from neural data and by developing methodological tools toward this goal. In particular, a method for localizing speech-related brain areas directly during surgery was needed in order to optimize the positioning of micro-electrode arrays (MEAs) that can only cover a limited surface of the brain.

In **Chapter 6**, we thus present a method to automatically map speech-related brain areas during awake brain surgery, in real-time. This method first automatically detects speech and extracts neural features which are then used to assess significant changes in the neural activity between silence and speech states.

In **Chapter 7**, we present preliminary results on speech intention detection, i.e. on decoding when a patient is speaking or intend to speak, on voicing activity detection, i.e. on decoding when the vocal folds are vibrating, and on decoding articulatory trajectories from neural data. To achieve this, we used data from two patients undergoing awake surgery for a tumor removal.

Chapter 6: Per-operative mapping of speech-related brain activity

I. Introduction

In the previous chapters, we presented an articulatory-based synthesizer which can be controlled in real-time by different speakers in order to produce intelligible speech. Such synthesizer could be used to synthesize speech using articulatory trajectories decoded from the neural activity of a subject's speech motor cortex.

During my thesis, we worked with Prof. Stephan Chabardès, who is a neurosurgeon at the university hospital (CHU) of Grenoble. This allowed us to record neural activity from patients undergoing awake brain surgery for a tumor resection (**Fig. 53**). Indeed, one treatment for brain tumors consists in removing the tumor areas and cells while preserving the sensory-motor functions of the patient. Localization of the functional areas to be preserved, such as the speech areas, is thus critical. It is often done by using neuroimaging techniques (e.g. fMRI), prior to surgery, but with limited spatial resolution. In some cases however, it is performed during the surgery by applying electrical stimulations at different locations on the cortex, while the patient is awake and self-reporting (Duffau et al., 2008, 2014). However, electrical stimulation gives only an indirect assessment of the functional areas, with uncontrolled spatial resolution, and may trigger epileptic seizures that are extremely problematic for open-brain surgery. Thus, another solution could be to perform the functional mapping of the cortical areas using electrophysiological recordings (Duffau et al., 2003; Boussen et al., 2016). In (Boussen et al., 2016), ECoG recordings were used to compare the spectral content of tumoural and healthy areas. Results showed that differences in the spectral content could allow tumor recognition directly during awake surgery. However, such approach does not give any information on the functional areas to be preserved.

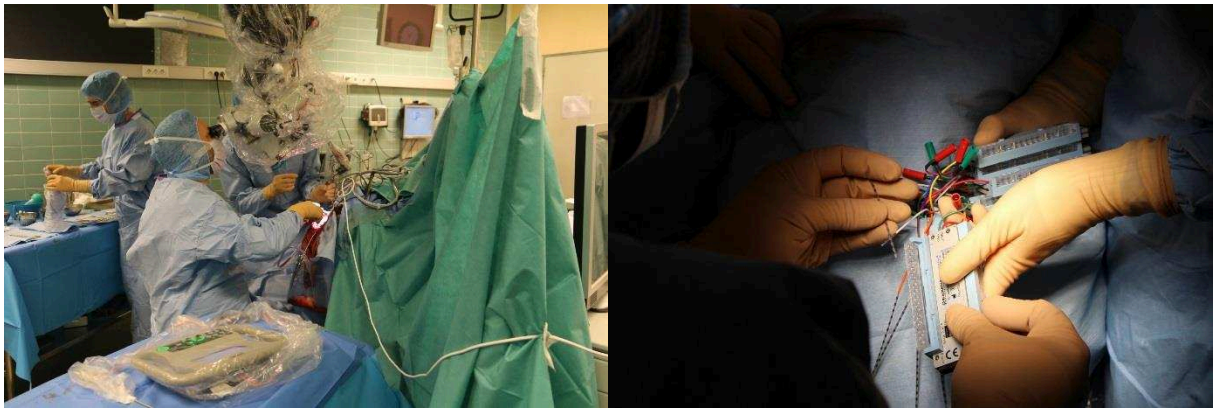


Fig. 53: Awake brain surgery at the hospital of Grenoble. Left – View of the operative room. Right – Connecting the 256 electrodes grid to the Blackrock system.

Moreover, the localization of speech-related brain areas during surgery could be of particular use to precisely position micro-electrode arrays (MEAs) that can only cover a limited cortical surface. Because each individual is different, and so is the fine organization of the brain, the best location for a MEA cannot be known in advance and has to be optimized for each

subject. While fMRI is a non-invasive technique that can provide some information on the candidate locations, its spatial resolution (typically, several millimeters) is rather limited compared to the size of an MEA. Thus, the mapping of speech-related brain areas could help finding the best positioning for an MEA in order to record speech related activity.

Here, we thus recorded neural activity using ECoG grids during awake surgeries in two patients for which the tumor was located next to the speech motor cortex. We also developed an approach to map speech-related brain activity, directly during awake brain surgery, and on the neurosurgeon's view of the operative field. Such mapping could allow to identify speech-related brain areas that should be preserved during the resection. This was also a good opportunity to record neural data during speech production in order to investigate the decoding of speech from neural data (see **Chapter 7**).

II. Methods

1. Subjects and experimental design

During my thesis, we worked with Prof. Stephan Chabardès, who is a neurosurgeon at the university hospital (CHU) of Grenoble. This allowed us to record neural activity from patients undergoing awake brain surgery for a tumor resection. Indeed, one treatment for brain tumors consists in removing the tumor areas and cells while preserving the sensory-motor functions of the patient. Localization of the functional areas to be preserved, such as the speech areas, is thus critical. It is often done by using neuroimaging techniques (e.g. fMRI), prior to surgery, but with limited spatial resolution. In some cases however, it is performed during the surgery by applying electrical stimulations at different locations on the cortex, while the patient is awake and self-reporting. However, electrical stimulation gives only an indirect assessment of the functional areas, with uncontrolled spatial resolution, and may trigger epileptic seizures that are extremely problematic for open-brain surgery. Thus, another solution could be to perform the functional mapping of the cortical areas using electrophysiological recordings. In that context, we were able to record neural activity using ECoG grid during awake surgeries in two patients for which the tumor was located next to the speech motor cortex.

a. First patient

The first patient (denoted P1 in the following) was a French male with no speech disorder, whose tumor was located in the upper part of the precentral gyrus. During the surgery, an ECoG grid (DIXI medical company, 10m spacing and 5mm diameter) was placed over the inferior prefrontal gyrus and its opercular part to record the neural activity of the patient while speaking overtly and covertly (**Fig. 54**-A). For this first surgery – and for regulatory reasons – we had to use the recording system in place at the hospital: the ISIS IOM system by InoMed (http://www.en.inomed.com), which could only record activity from 4 electrodes of the ECoG grid, as shown on **Fig. 54**-B. A microphone was fixed near the patient to record his voice during the whole experiment. Since the InoMed system cannot record any other external signal than the brain signals, we used a CED Micro1041 data acquisition unit (http://ced.co.uk) to record

the audio signal as well as a trigger signal used for synchronization purpose. This trigger signal was also sent to the InoMed system in order to trigger short recordings of the neural data generally used for evoked potentials recordings. The times of each trigger in the neural data could be recovered by comparing each of these small recording with the whole neural signal. This trick was used to synchronize the neural data recorded by the InoMed system, and the audio signal recorded by the CED Micro1041. The neural data was recorded at 2kHz while the audio signal was recorded at 50kHz.

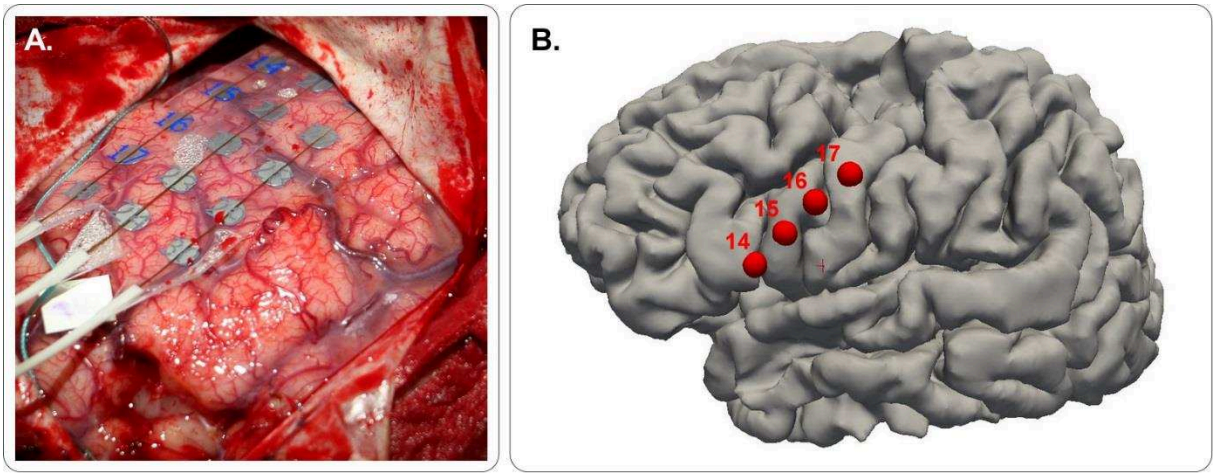


Fig. 54: Positionning of the ECoG grid for the first patient. A – picture of the ECoG grid taken during the surgery. Only the electrodes which number is in blue (14 to 17) were recorded. B – localization of the 4 recorded electrodes on a reconstruction of the brain geometry from IRM data using the FreeSurfer software (freesurfer.net). The electrodes were localized by pointing them with the navigation tool of the surgery room. The numbers of each electrode are those of the electrodes in B.

The patient was first asked to overtly pronounce (i.e. saying out loud) isolated vowels or vowel-consonant-vowel sequences (VCVs). More precisely, he had to pronounce the 10 vowels /a/, /i/, /u/, /o/, /œ/, /e/, /y/, /ã/, /ɛ̃/, and /ɔ̃/, and the 30 VCVs made of /b/, /d/, /g/, /v/, /z/, /ʒ/, /m/, /n/, /r/, and /l/, in /a/, /i/ and /u/ contexts (i.e., ‘aba’, ‘idi’, ‘umu’, and so on). The patient was asked to repeat three times in a row each item after hearing the corresponding audio extracted from the PB2007 dataset, played through speakers that were placed in front of him. On average, each item was pronounced 6 times by the patient. Because of the noisy hospital environment, and the overall tiredness of the patient, somes items were mistaken with others, thus resulting in more or less repetitions per item.

The patient was then asked to covertly pronounce (i.e. to imagine pronouncing without actually moving or producing any sound) the 3 vowels /a/, /i/ and /u/, as well as the 9 VCVs made of /b/, /d/ and /g/ in /a/, /i/ and /u/ vocalic contexts. The audio of each item – from the PB2007 dataset - was first played to the patient three times at a fixed pace, after what he had to continue to imagine repeating them three times while keeping the same pace. Playing each item three times ensured that the patient could catch and keep the same pace, which was crucial to estimate when he was atually covertly speaking. That procedure was repeated twice, so that theoretically the patient imagined each item 6 times in total.

The whole experimental protocol, including the setting up of the reording system, had to fit in a 30 minutes time window during the surgery in order to not prolongate too much the surgery. For this first patient, the localization of the speech-related brain areas was performed after the experiment. This was done using our custom made software, called NeuroPXI (Bonnet et al., 2012), which can replay a previously recorded file in order to simulate a real-time

experiment (**Fig. 55**). During this thesis, I had the chance to supervise an intern, Eloi Navarro, who implemented this specific functionality in the NeuroPXI software.

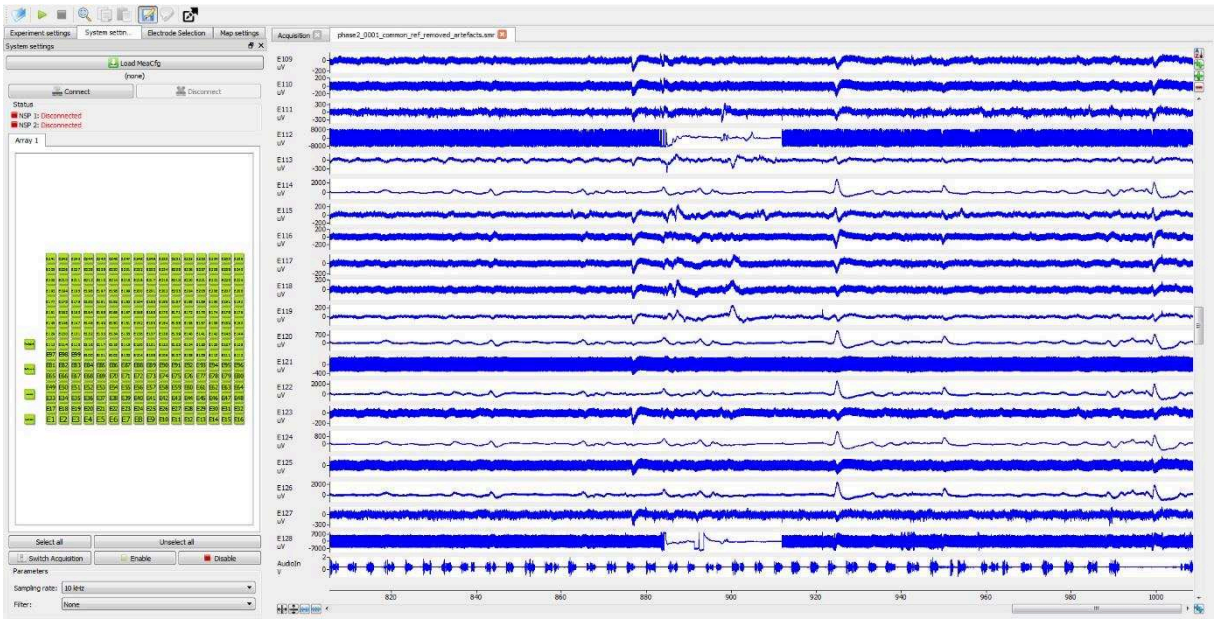


Fig. 55: The NeuroPXI software. This software allows to stream the recorded neural data in real-time, as well as to replay previously recorded files. Some channels were excluded from the analysis because of their noise level (e.g. second channel from the bottom).

b. Second patient

The second patient (denoted P2 in the following) was a French male with no speech disorder, whose tumor was located in the left central sulcus, between the motor and sensory cortices (see **Fig. 56**-A). During the experiment, a custom made 256 electrodes ECoG grid (PMT corporation company, 3mm vertical and 3.5mm horizontal spacings, 1mm diameter) was placed over the exposed area (**Fig. 56**-B), covering some parts of the speech motor and sensory cortices (**Fig. 56**-A). The neural activity was recorded at 10 kHz using two synchronized 128-channel Blackrock Microsystems recording units (http://blackrockmicro.com/). A directionnal microphone was placed in front of the patient, and speech was recorded at 10 kHz using an external analog input, thus ensuring perfect synchronization with the neural signal. Using a directionnal microphone allowed to have a good quality recording of the patient voice, while removing most of the environmental noise, surgery rooms being particularly noisy. The data was recorded using our NeuroPXI software which was extended in order to support recordings from the Blackrock system.

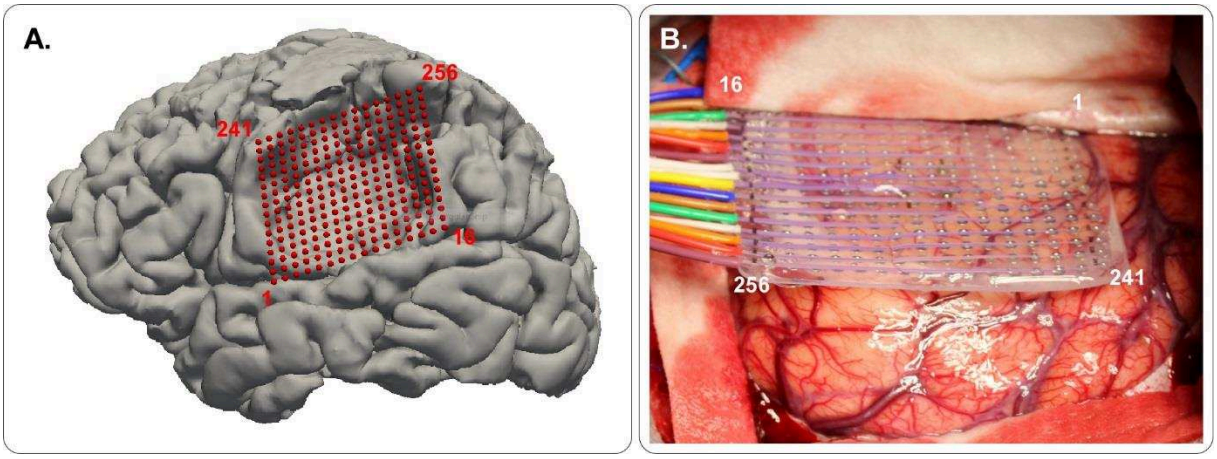


Fig. 56: Positionning of the ECoG grid for the second patient. A – Approximate localization of the 256 recorded electrodes on a reconstruction of the brain geometry from IRM data using the FreeSurfer software (freesurfer.net). The electrodes were localized by pointing some of them with the navigation tool of the surgery room, and then interpolating the coordinates for the other electrodes. The numbers of the electrodes at the extremities of the grid are the same as in B. B – Picture of the ECoG grid taken during the surgery.

The patient was first asked to perform continous movement of each of the main speech articulators: switching from a kiss to an exagerated smile with the lips; moving the jaw up and down by opening and closing the mouth widely; moving the tongue back and forth; changing the larynx activity by alternating the sound /s/ with its voiced counterpart /z/; and moving the velum by alternating the sound /o/ with its nasal counterpart /õ/.

The patient was then asked to pronounce 3 times each of the 10 French vowels /a/, /i/, /u/, /o/, /œ/, /e/, /y/, /ã/, /ɛ̃/, and /õ/, and then to read out loud a list of short sentences (about 4-5 words each, extracted from the BY2014 corpus), at the pace he felt more comfortable with. Each sentence was displayed at the center of a tablet screen, and one experimenter was in charge of passing to the next sentence once the patient was done pronouncing the currently displayed sentence. In total, the patient pronounced about two hundred sentences.

The whole experimental protocol, including connecting the electrodes and setting up the recording system, had to last about 60 minutes during the surgery in order not to prolongate too much the surgery. For this second patient, and for both tasks, the neural data was processed online in order to display the speech-related brain activity directly on a picture of the operative field taken before placing the ECoG grid, in real-time. This required to automatically detect speech instants using the microphone input, then to combine this information with features extracted from the neural signal in order to identify how these features differed between speech and silent periods, and finally to map the significant speech-related brain activity on a picture of the brain using registration techniques. Each of these processing parts are described in the following.

2. Automatic speech detection

In order to localize speech-related brain areas, one must first be able to identify which segments of data correspond to speech, and which correspond to silence. This can be achieved by automatically analyzing the audio signal coming from the microphone placed next to the patient. In the present work, the good quality of the recorded audio signals allowed us to use a simple threshold approach (**Fig. 57**).

During speech, the audio signal energy is generally higher. Thus, the algorithm we used ensured that audio samples which absolute value was above a fixed threshold were labeled as speech samples and samples below the threshold as silence samples (**Fig. 57**-B). However, speech is time-varying signal, so that during speech, the audio signal absolute value will go several times above and below the threshold (**Fig. 57**-B, inset). To account for that, all samples in-between two speech samples were as well considered as speech if the two speech samples were close enough, i.e. if they were separated by less than some time period, called the inactivation period (**Fig. 57**-C). Choosing this inactivation period long enough (here, 300ms) allowed as well to consider short pauses within sentences or sounds with low energy (e.g. /s/) as speech segments **Fig. 57**-D). Moreover, in order to reject short environmental noises, such as the beeps emitted by the equipments of the surgery room, all speech segments that were shorter than a fixed duration (here, 500ms) were considered as non-speech segments, i.e. as silence.

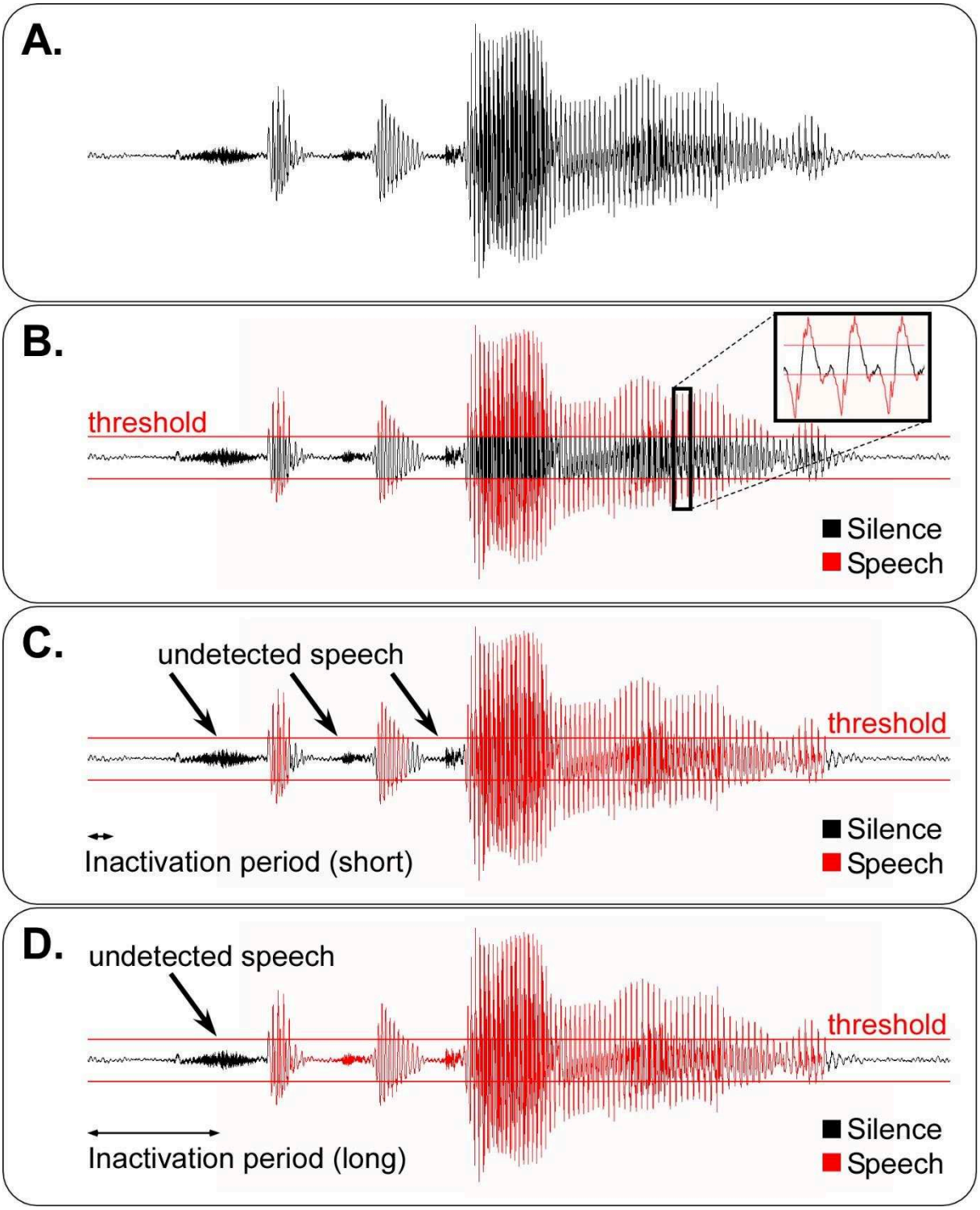


Fig. 57: Automatic online speech detection. A – Raw audio signal recorded by a microphone placed next to the patient. B – Samples which absolute value is above threshold are labeled as speech (red), and samples below threshold are labeled as silence (black). The speech signal is a time-varying signal, thus many speech samples are not detected (inset). C – Using an inactivation period allows to include fast oscillation in the speech signal. However, short pauses and phones with low energy remain undetected (arrows). D – Using a larger inactivation period allows to include short pauses and low-energy phones that are in-between high-energy phones. However, low-energy phones remain undetected at the beginning of the speech signal (black arrow).

The parameters of the automatic speech detection were optimized prior to the beginning of the experiment by asking the subject to speak while adjusting the dection parameters. Once

speech and non-speech segments were discriminated, speech-related brain activity could be detected by analyzing differences in the neural signal between these segments.

3. Extraction of speech-related brain activity

One common way to extract information from ECoG neural signals is to decompose it into frequency components and compute its power at each frequency – i.e. its spectrum – or band of frequencies, at different time instants. Several ECoG studies have shown that cortical oscillations are relevant correlates of speech processing (Leuthardt et al., 2011; Pei et al., 2011a; Pasley et al., 2012; Bouchard et al., 2013; Pasley and Knight, 2013; Martin et al., 2014; Mugler et al., 2014). In particular, speech production is classically associated with a decrease of signal power in the beta frequency range (10-30Hz) and usually an increase in the high-gamma frequency range (70-200Hz) over temporal and frontal areas (Canolty et al., 2007; Pei et al., 2011b; Toyoda et al., 2014) while gamma attenuation was observed in more anterior frontal speech cortex including Broca's area (Lachaux et al., 2008; Wu et al., 2011; Toyoda et al., 2014). These oscillatory features can thus be used to map functional cortical speech areas during resection surgeries (Kamada et al., 2014; Tamura et al., 2016). Speech-related brain areas were thus identified by quantifying differences in spectral power between speech and silence segments.

For each recorded electrode of the ECoG grid, the short term Fourier transform (STFT) of the neural signal was computed in real-time in order to obtain the instantaneous power for each frequency. The STFT was performed using a 512 samples (i.e. 256ms at 2kHz) window for patient P1, and 2048 samples (i.e. 205ms at 10kHz) for P2, with 75% of overlap between consecutives windows, and after windowing the data using a Hamming function.

Using the output from the automatic speech detection, the STFT was then averaged over time to estimate, for each electrode i, the mean power of the neural signal for the frequency f during silence $\mu^i_{silence}(f)$, and the one during speech $\mu^i_{speech}(f)$ along with their respective standard deviations $\sigma^i_{silence}(f)$ and $\sigma^i_{speech}(f)$. Here, we assumed that the power $P^i(f)$ of the neural signal of each electrode i and frequency f had a Gaussian distribution during speech and silence:

$$P^i_{silence}(f) \sim N(\mu^i_{silence}(f), \sigma^i_{silence}(f)) \quad \text{Eq. 28}$$

$$P^i_{speech}(f) \sim N(\mu^i_{speech}(f), \sigma^i_{speech}(f)) \quad \text{Eq. 29}$$

Under this assumption, significant changes in the neural activity power between speech and silence could be identified using a two-tailed Welch's t-test, for each electrode i and frequency f. The Welch's t-test is then performed by first computing the statistics $t^i(f)$ and the number of degrees of freedom $v^i(f)$:

$$t^i(f) = \frac{\mu^i_{speech}(f) - \mu^i_{silence}(f)}{\sqrt{\frac{\sigma^i_{speech}(f)^2}{N_{speech}} + \frac{\sigma^i_{silence}(f)^2}{N_{silence}}}} \quad \text{Eq. 30}$$

$$v^i(f) = \left\lfloor \frac{\left(\frac{\sigma^i_{speech}(f)^2}{N_{speech}} + \frac{\sigma^i_{silence}(f)^2}{N_{silence}}\right)^2}{\frac{\left(\frac{\sigma^i_{speech}(f)^2}{N_{speech}}\right)^2}{N_{speech}+1} + \frac{\left(\frac{\sigma^i_{silence}(f)^2}{N_{silence}}\right)^2}{N_{silence}+1}} \right\rfloor - 2 \quad \text{Eq. 31}$$

Where N_{speech} and $N_{silence}$ are the number of averaged STFT used to obtain the estimate mean and standard deviation for speech and silence, respectively. The P-value $P^i(f)$ for this electrode i and frequency f can then be computed using:

$$P^i(f) = 2 * \tilde{T}_{v^i(f)}(t^i(f)) \quad \text{Eq. 32}$$

Where $\tilde{T}_v$ is the cumulative distribution function of the complement of the Student distribution with v degrees of freedom. Since this P-value is computed for each electrode and frequency, the Bonferroni correction for comparison was applied so that $P^i(f)$ is significant only if $P^i(f) < \tilde{\alpha}$, with $\tilde{\alpha}$ the Bonferroni corrected risk factor (Bonferroni, 1936) given by:

$$\tilde{\alpha} = \frac{\alpha}{N_{elec} * N_f} \quad \text{Eq. 33}$$

With α the original risk factor, N_{elec} the number of electrodes and N_f the number of frequencies. In the following, α was chosen equal to 0.05, and the corrected risk factor was then computed for each patient according to the number of electrodes and frequencies considered.

These significant changes in neural activity could then be quantified by computing $R^i(f)$, which will be called the speech-silence-ratio in the following, as well as the z-score $Z^i(f)$:

$$R^i(f) = \frac{\mu^i_{speech}(f)}{\mu^i_{silence}(f)} - 1 \quad \text{Eq. 34}$$

$$Z^i(f) = \frac{\mu^i_{speech}(f) - \mu^i_{silence}(f)}{\sigma^i_{silence}(f)} \quad \text{Eq. 35}$$

That way we obtained, for each electrode i and each frequency f a speech-silence-ratio which represented the degree of difference in the neural activity during speech as compared to silence for this specific frequency. A positive speech-silence-ratio represented an increase in activity during speech while a negative value represented a suppression of the activity during speech.

Finally, frequencies of interest could be identified by computing $C(f)$, the number of electrodes that had significant difference of activity between the speech and silence conditions, given a risk factor α:

$$C(f) = \sum_{i=1}^{N} \delta_i(f) \text{ with } \delta_i(f) = \begin{cases} 1, if\ P_i(f) < \tilde{\alpha} \\ 0, otherwise \end{cases} \quad \text{Eq. 36}$$

Frequencies with the highest C values corresponded to frequencies at which many electrodes exhibited speech-specific activity. The visualization of speech activity was done by

directly mapping the speech-silence-ratio at the electrode positions, for each frequency of interest, when it was significant.

4. Mapping of speech-related brain activity

When using an electrode grid, the neural activity, and all derived features such as the speech-score presented in the previous part, are sampled only at discrete spatial positions corresponding to the recording sites. Therefore, in order to visualize the brain activity over the whole covered area, an interpolation method is required to extend the available data to all the positions between electrodes. Here we used thin plate splines – or surface splines – interpolation (Perrin et al., 1987), which was previously used in another neural activity mapping software developped by our team: the NeuroMap software (Abdoun et al., 2011), which is publicly available (https://sites.google.com/site/neuromapsoftware/).

Compared to conventional interpolation methods, such as bilinear or bicubic interpolation, thin plate spline interpolation has several advantages: (1) it allows the estimation of activity with local extrema not necessarily at recording sites unlike bilinear interpolation, (2) it does not require equally spaced spatial sampling and can be thus used for any disposition of electrodes, (3) it can be computed efficiently and (4) it provides an analytical expression of the interpolant, which is differentiable.

If we consider n recording electrodes, and note (x_i, y_i) the coordinates of the electrode i and v_i the measured value at this electrode, then its interpolated value v_m at the coordinates (x, y) using a m[th] order surface spline is given by:

$$v_m(x,y) = \sum_{i=1}^{n} p_i k_m(x - x_i, y - y_i) + \sum_{d=0}^{m-1} \sum_{k=0}^{d} q_{kd} x^{d-k} y^k \qquad \text{Eq. 37}$$

Where k_m is the function defined by:

$$k_m(x,y) = (s^2 + t^2)^{m-1} \log(s^2 + t^2) \qquad \text{Eq. 38}$$

The n coefficients p_i and the $\frac{m(m+1)}{2}$ coefficients q_{kd} can be obtained through the resolution of a system of linear equation. By noting $\underline{\boldsymbol{V}} = (v_i)_{1 \le i \le n}$, $\underline{\boldsymbol{P}} = (p_i)_{1 \le i \le n}$ and $\underline{\boldsymbol{Q}} = \left(q_{00}, q_{01}, q_{11}, \dots, q_{m-1,m-1}\right)^T$, the coefficients of $\underline{\boldsymbol{P}}$ and $\underline{\boldsymbol{Q}}$ can be obtained by solving the following system of linear equations:

$$\begin{bmatrix} \underline{\boldsymbol{K}} & \underline{\boldsymbol{E}} \\ \underline{\boldsymbol{E}^T} & \underline{\boldsymbol{0}} \end{bmatrix} \begin{bmatrix} \underline{\boldsymbol{P}} \\ \underline{\boldsymbol{Q}} \end{bmatrix} = \begin{bmatrix} \underline{\boldsymbol{V}} \\ \underline{\boldsymbol{0}} \end{bmatrix} \qquad \text{Eq. 39}$$

Where

$$\underline{\boldsymbol{K}} = \left(k_{ij}\right) with\ k_{ij} = k_m(x_i - x_j, y_i - y_j) \qquad \text{Eq. 40}$$

$$\underline{\boldsymbol{E}} = \begin{bmatrix} 1 & x_1 & y_1 & x_1^2 & x_1 y_1 & \cdots & x_1 y_1^{m-2} & y_1^{m-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_n & y_n & x_n^2 & x_n y_n & \cdots & x_n y_n^{m-2} & y_n^{m-1} \end{bmatrix} \qquad \text{Eq. 41}$$

And which can be rewritten as $\underline{\boldsymbol{AX}} = \underline{\boldsymbol{B}}$ with $\underline{\boldsymbol{A}} = \begin{bmatrix} \underline{\boldsymbol{K}} & \underline{\boldsymbol{E}} \\ \underline{\boldsymbol{E}^T} & \underline{\boldsymbol{0}} \end{bmatrix}$, $\underline{\boldsymbol{X}} = \begin{bmatrix} \underline{\boldsymbol{P}} \\ \underline{\boldsymbol{Q}} \end{bmatrix}$ and $\underline{\boldsymbol{B}} = \begin{bmatrix} \underline{\boldsymbol{V}} \\ \underline{\boldsymbol{0}} \end{bmatrix}$. A numerical approximate solution $\underline{\tilde{\boldsymbol{X}}}$ of this system is:

$$\underline{\tilde{\boldsymbol{X}}} = \underline{\boldsymbol{A}}^{+}\underline{\boldsymbol{B}} \qquad \text{Eq. 42}$$

Where $\underline{\boldsymbol{A}}^{+}$ is the Moore-Penrose pseudo-inverse of $\underline{\boldsymbol{A}}$, which can be obtained using the singular value decomposition of $\underline{\boldsymbol{A}}$:

$$\underline{\boldsymbol{A}} = \underline{\boldsymbol{USV}}^T \Longrightarrow \underline{\boldsymbol{A}}^{+} = \underline{\boldsymbol{V}}\underline{\tilde{\boldsymbol{S}}}^{+}\underline{\boldsymbol{U}}^T \qquad \text{Eq. 43}$$

With $\underline{\boldsymbol{S}} = (s_i)_{1 \le i \le 2n}$ a diagonal matrix containing the singular values of A from the highest to the lowest, so that $\underline{\tilde{\boldsymbol{S}}}^{+} = (\tilde{s}_i^{+})_{1 \le i \le 2n}$ with:

$$\tilde{s}_i^{+} = \begin{cases} \dfrac{1}{s_i}, if\ i \le k \\ 0, if\ i > k \end{cases} \qquad \text{Eq. 44}$$

In most cases, k is chosen equal to the number of non-zero values of S. However, a smaller value of k can be considered when $\underline{\boldsymbol{A}}$ is ill-conditioned in order to obtain a regularized solution (Uutela et al., 1999).

The matrices $\underline{\boldsymbol{P}}$ and $\underline{\boldsymbol{Q}}$ can then be obtained using the following:

$$\begin{bmatrix} \underline{\boldsymbol{P}} \\ \underline{\boldsymbol{Q}} \end{bmatrix} = \underline{\boldsymbol{A}}^{+} \begin{bmatrix} \underline{\boldsymbol{V}} \\ \underline{\boldsymbol{0}} \end{bmatrix} \qquad \text{Eq. 45}$$

Thus, if the source positions of the interpolant do not change – which was the case here since they are electrodes coordinates – the matrix $\underline{\boldsymbol{A}}^{+}$ only needs to be computed once even if the measured values change. In a similar manner, if the interpolated positions do not change – which was the case here since they are all the pixel positions in between electrodes – the interpolated values can be efficiently obtained by pre-computing and storing all the k_m function values at the interpolation points (x, y) as well as the $x^{d-k}y^{k}$ terms in the analytical expression of the interpolant. To further reduce the computational complexity of this method, the area covered by the ECoG grid was subdivided into a dense NxM point grid. Values where interpolated using thin plate spline interpolation at each point of this grid, and points in-between where then further interpolated using bilinear interpolation. Indeed, a computer screen has about two million visible pixels. When displaying the interpolated map full-screen, it would require to compute about the same amount of interpolated values, which would not be efficient. Using a subdisivion grid resulted in computing less interpolated values while keeping a good spatial resolution. For instance, for a typical 10cm x 10cm ECoG grid, using a 100x100 subdivision grid would result in computing only 10,000 interpolated values while keeping a spatial resolution of 1mm with spline interpolation.

Using such interpolation technique allowed to map the speech-silence-ratio over the whole brain surface covered by the ECoG grid, in real-time. The P-values given by the Welch's t-test were as well interpolated in order to prevent unsignificant differences to be displayed. In order to clearly identify which brain areas exhibited speech-specific activity, localization of the

electrode positions on anatomical pictures was needed, and was done on a picture of the brain directly taken during the surgery.

5. Coregistration of the electrodes on the operative field

Mapping the neural activity directly on the surgeon's view of the operative field is of particular interest in order to identify precisely the best location for a micro-electrode array to record speech-related activity, and is also useful to inform the surgeon of which areas are to be preserved during the resection. If we considered mapping the neural activity on the brain volume reconstructed from the MRI data, navigation tools from the surgery room could have been used to directly obtain the coordinates of each electrode in the MRI data by pointing them with the surgery navigation device. Indeed, during the surgery the patient was fixed in a stereotaxic frame which allowed to know the localization of the patient within the MRI space. However, we considered directly mapping the neural activity on the surgeon's view – here on a picture of the exposed brain taken during the surgery – so that this information was not available and another approach was needed.

We used a similar techniques to that of the approach implemented in the NeuroMap software (https://sites.google.com/site/neuromapsoftware/). Two pictures of the exposed brain were taken during the surgery: one before placing the ECoG grid, so that the full anatomy was visible, and one after placing the ECoG grid, so that the electrodes were visible but masking relevant anatomical parts. In a first step, the electrodes were localized on the second picture in which electrodes were visible. In a second step, correspondence points between the two pictures – with and without electrodes – were identified so that the electrode positions on the first picture could be inferred from their known positions in the second picture (**Fig. 58**).

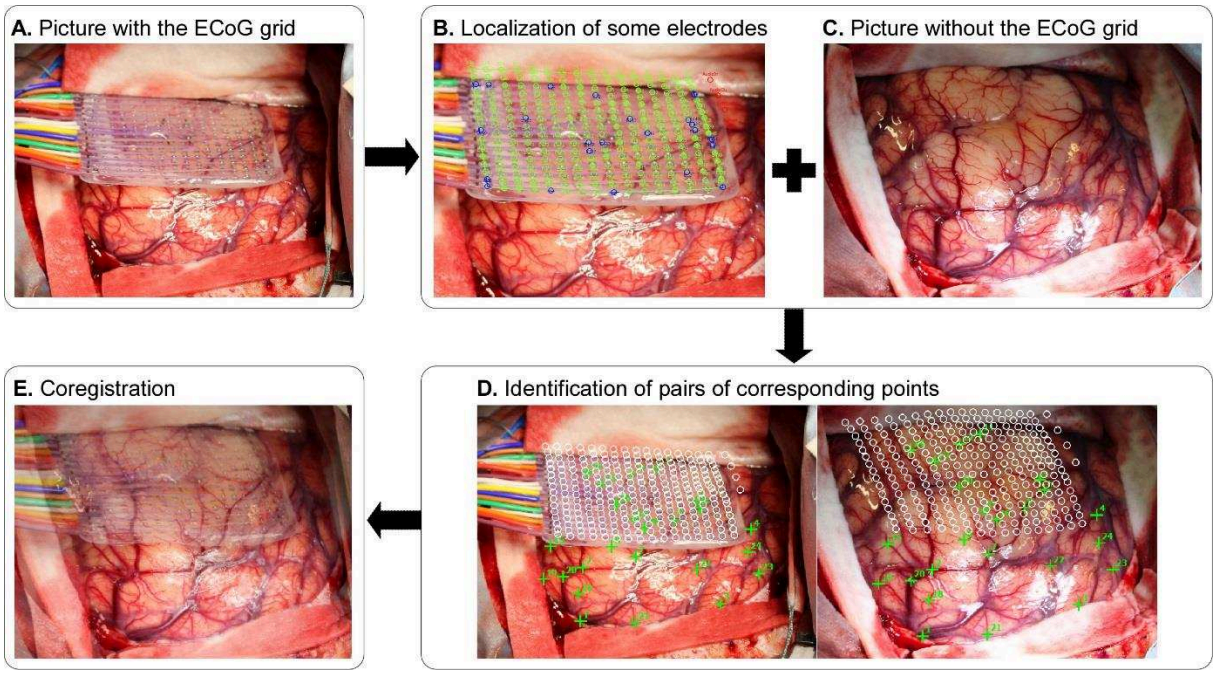


Fig. 58: Coregistration of the electrodes on the anatomy. A - During the surgery a picture of the exposed brain with the ECoG grid is taken. B - Some electrodes are localized on the picture by the user (blue circles), which allow to infer the position of all the other electrodes (green circles). C - Before placing the ECoG grid on the brain, a picture of the exposed brain without the grid was taken. D - Pairs of corresponding points between both pictures are identified by the user (green crosses, corresponding points are labeled by an identical number), which allows to infer the position of the electrodes on the anatomy

(white circles on the right image) using their positions on the picture with the ECoG grid visible (white circles on left picture). E - The quality of the coregitration can be evaluated by superimposing the picture without the ECoG grid with a deformed version of the picture with the ECoG grid using the same thin plate spline interpolation that inferred the electrodes positions. Here we see that it allows to recover the anatomy below the ECoG grid and that there is no significant mismatch (otherwise we would observe some vessel or other anatomical features twice, resulting in a blurry image).

For the first step, this method takes advantage of knowing in advance the geometry of the ECoG grid. Indeed, ECoG grids were designed long before the surgery, to that the position of each electrode in the grid space was known. The user started by identifying the positions of n electrodes, by simply pointing them in the picture with the ECoG grid visible (**Fig. 58**-A), allowing to obtain each of their pixel coordinates $(x_{i_{pixel}}, y_{i_{pixel}})$. Using the known geometry of the ECoG grid, the algorithm then used the grid coordinates $(x_{i_{grid}}, y_{i_{grid}})$ to extrapolate the pixel coordinates $(x_{j_{pixel}}, y_{j_{pixel}})$ of all the N electrodes of the grid $(1 \leq j \leq N)$ using a second order thin plate spline interpolation, as previously defined:

$$x_{j_{pixel}} = \sum_{i=1}^{n} p_i^{(X)} k_2\left(x_{j_{grid}} - x_{i_{grid}}, y_{j_{grid}} - y_{i_{grid}}\right) + \sum_{d=0}^{1}\sum_{k=0}^{d} q_{kd}^{(X)} x_{j_{grid}}^{d-k} y_{j_{grid}}^{k} \quad \text{Eq. 46}$$

$$y_{j_{pixel}} = \sum_{i=1}^{n} p_i^{(Y)} k_2\left(x_{j_{grid}} - x_{i_{grid}}, y_{j_{grid}} - y_{i_{grid}}\right) + \sum_{d=0}^{1}\sum_{k=0}^{d} q_{kd}^{(Y)} x_{j_{grid}}^{d-k} y_{j_{grid}}^{k} \quad \text{Eq. 47}$$

$$\underline{\boldsymbol{V}^{(X)}} = \begin{bmatrix} x_{1_{pixel}} \\ \vdots \\ \boldsymbol{x_{n_{pixel}}} \end{bmatrix} \text{ and } \underline{\boldsymbol{V}^{(Y)}} = \begin{bmatrix} y_{1_{pixel}} \\ \vdots \\ \boldsymbol{y_{n_{pixel}}} \end{bmatrix} \quad \text{Eq. 48}$$

Note that in that case, the linear system has to be solved twice, for $\underline{\boldsymbol{V}^{(X)}}$ then for $\underline{\boldsymbol{V}^{(Y)}}$, in order to obtain the matrices $\underline{\boldsymbol{P}^{(X)}}$, $\underline{\boldsymbol{Q}^{(X)}}$, $\underline{\boldsymbol{P}^{(Y)}}$, and $\underline{\boldsymbol{Q}^{(Y)}}$. However, the singular decomposition of the matrix $\underline{\boldsymbol{A}}$ could be done once, as well as the computation of the k_2 coefficients and $x_{j_{grid}}^{d-k} y_{j_{grid}}^{k}$ terms, which saved computational power.

Thus, the precision of the electrode localization on the picture increases with the number of electrodes identified by the user. In practice, it was often sufficient to identify about ten electrodes (blue dots in **Fig. 58**-B), which saved a lot of experimental time, especially when the ECoG grid contained more than two hundred electrodes. Thus, at the end of this step, the position of all electrodes in the pictures with the ECoG grid visible were known (**Fig. 58**-B).

In the second step, these positions were used to infer the positions of the electrodes in the picture without the ECoG grid (**Fig. 58**-C). This required to coregister both pictures. Indeed, the pictures could have been taken using a different angle and camera setting. If we note $(x_{j_{pixel}}^{(1)}, y_{j_{pixel}}^{(1)})$ the pixel coordinates of the ith electrodes in the picture with the ECoG grid visible, then the aim of the coregistration algorithm was to provide $(x_{j_{pixel}}^{(2)}, y_{j_{pixel}}^{(2)})$, the pixel coordinates of the ith electrode in the picture without the ECoG grid, for each of the N electrodes of the grid. This was achieved by first identifying n anatomical landmarks in both pictures $\{(\hat{x}_{i_{pixel}}^{(1)}, \hat{y}_{i_{pixel}}^{(1)}); (\hat{x}_{i_{pixel}}^{(2)}, \hat{y}_{i_{pixel}}^{(2)})\}$, for instance the crossing of some blood vessels that were visible in both pictures, which were then used to obtain the coordinates of all the electrodes using thin plate spline interpolation similarly to the previous step (**Fig. 58**-D):

$$x_{j_{pixel}}^{(2)} = \sum_{i=1}^{n} p_i^{(X)} k_2\left(x_{j_{pixel}}^{(1)} - \hat{x}_{i_{pixel}}^{(1)}, y_{j_{pixel}}^{(1)} - \hat{y}_{i_{pixel}}^{(1)}\right) + \sum_{d=0}^{1}\sum_{k=0}^{d} q_{kd}^{(X)} (x_{j_{pixel}}^{(1)})^{d-k} (y_{j_{grid}}^{(1)})^{k} \quad \text{Eq. 49}$$

$$y_{j_{pixel}}^{(2)} = \sum_{i=1}^{n} p_i^{(Y)} k_2\left(x_{j_{pixel}}^{(1)} - \hat{x}_{i_{pixel}}^{(1)}, y_{j_{pixel}}^{(1)} - \hat{y}_{i_{pixel}}^{(1)}\right) + \sum_{d=0}^{1}\sum_{k=0}^{d} q_{kd}^{(Y)} (x_{j_{pixel}}^{(1)})^{d-k} (y_{j_{pixel}}^{(1)})^{k} \quad \text{Eq. 50}$$

$$\underline{\boldsymbol{V}^{(X)}} = \begin{bmatrix} \hat{x}_{1_{pixel}}^{(2)} \\ \vdots \\ \hat{x}_{n_{pixel}}^{(2)} \end{bmatrix} \text{ and } \underline{\boldsymbol{V}^{(Y)}} = \begin{bmatrix} \hat{y}_{1_{pixel}}^{(2)} \\ \vdots \\ \hat{y}_{n_{pixel}}^{(2)} \end{bmatrix} \quad \text{Eq. 51}$$

Thus, the more landmarks were identified, the more precise was the coregistration. In practice, a dozen of pairs was often enough to obtain a good registration of the pictures. Deforming the first picture in order for it to match the second one could allow to use the electrodes pixel coordinates in the second picture as coordinates in the first one. Here however, we chose to infer the electrode positions in the first picture from their positions in the second one so that the anatomical picture was not distorted, which could have made the localization more difficult or less accurate. However, we still used such image deformation in order to superimpose both pictures to qualitatively evaluate the quality of the coregistration while adding pairs of corresponding points (**Fig. 58**-E).

Using this coregistration approach thus allowed to identify the positions of the ECoG grid electrodes on the anatomy of the brain using a picture directly taken during the surgery, as well as on rendered images of the brain volume obtained from MRI prior to surgery. These positions could then be used to directly map the speech-silence-ratio on the surgical view in real-time, allowing to identify candidate areas to place a micro-electrode array for recording speech-specific brain activity.

6. Coregistration of the electrodes on the reconstructed cortical surface

In order to localize the electrode locations relatively to the different brain regions, we also estimated their positions on the reconstructed cortical surface of each subject. Each patient underwent magnetic-resonance imaging (MRI) prior to surgery in order to precisely localize the tumor to be removed. We used this MRI to reconstruct a 3D mesh of the cortical surface (**Fig. 54**-B and **Fig. 56**-A) using the FreeSurfer software (freesurfer.net).

During the surgery, the neurosurgeon pointed three anatomical landmarks using the neuronavigation system of the surgery room (**Fig. 59**, top row). Unfortunately, the neuronavigation software only displayed the MRI slices corresponding to the pointed location, and not its actual coordinates in the MRI spaces. In order to find these coordinates, we thus captured the displayed MRI slices (**Fig. 59**, middle row) that were later identified by exploring the full MRI data. This was performed using the 3DSlicer software (slicer.org) and allowed to obtain the coordinates $(R_n; A_n; S_n)$ of each pointed location n $(1 \leq n \leq 3)$ in the MRI space (**Fig. 59**, bottom row).

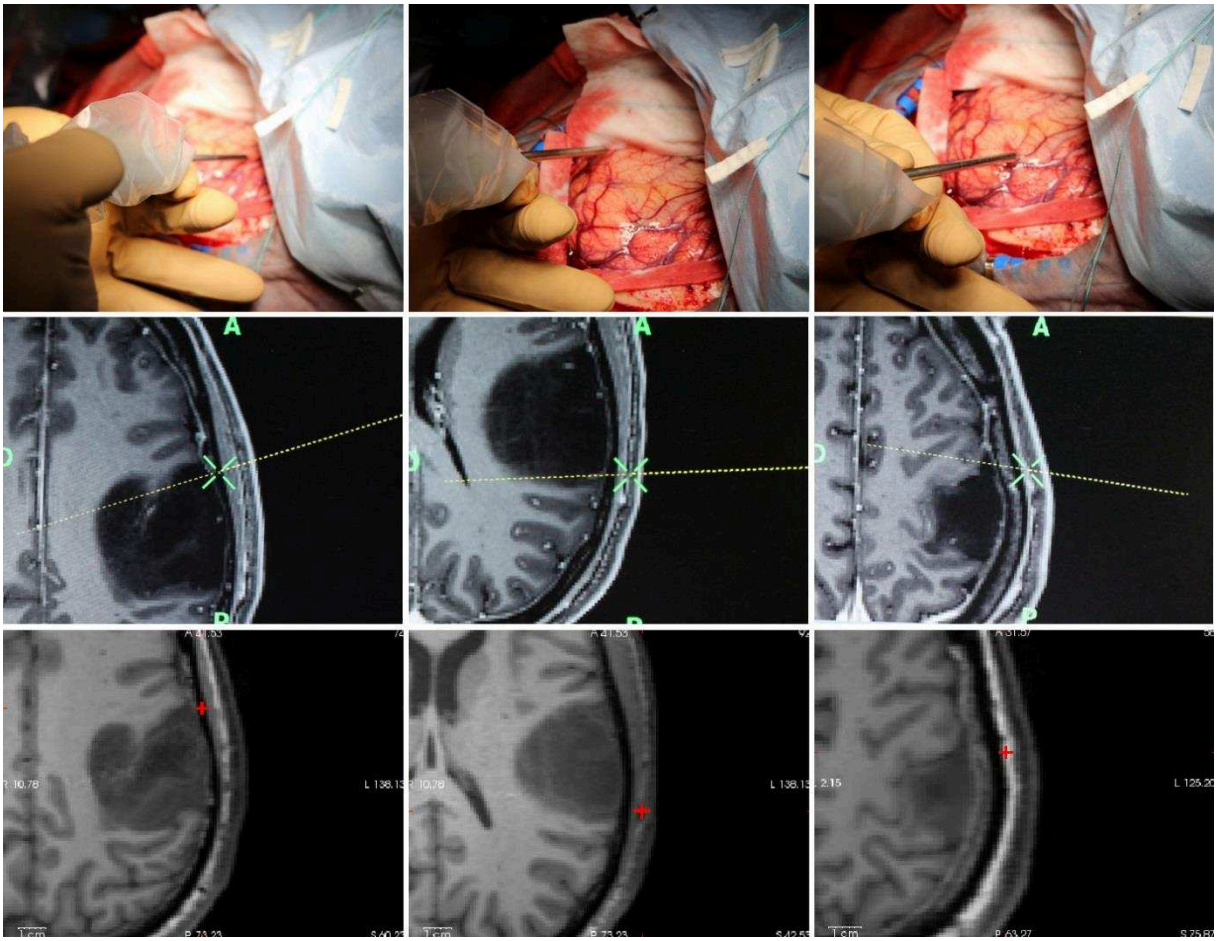


Fig. 59: Localization of anatomical landmarks in the MRI data. Top row – The three different anatomical landmarks pointed by the neurosurgeon. Middle row – The captured view of the neuronavigation system in the horizontal plane. The green cross indicates the localization of the pointed landmark. Bottom row – Corresponding location (red cross) manually identified using the 3DSlicer software.

The anatomical landmarks were as well identified in the picture of the operative field without the electrodes visible (**Fig. 58**-C). We then transformed the position of each anatomical landmark n in the picture to coordinates (i_n, j_n) in the coordinate space of the electrode grid (for instance, the coordinates (0;0) correspond to the electrode in the first row and first column, while the coordinates (1;1) correspond to the electrode in the second row and second column). This was achieved by using the thin-plate splines interpolation computed for the coregistration of the electrodes on the operative field (see previous section). By assuming that all electrodes are contained in the same 3D plane (i.e. by neglectin the curvature of the electrode grid), we thus obtain a direct linear relationship between the coordinates $(R; A; S)$ in the MRI space and the coordinates $(i; j)$ in the grid space:

$$\begin{cases} R(i,j) = R^{ref} + i * R^x + j * R^y \\ A(i,j) = A^{ref} + i * A^x + j * A^y \\ S(i,j) = S^{ref} + i * S^x + j * S^y \end{cases} \quad \text{Eq. 52}$$

Where R^{ref}, R^x, R^y, A^{ref}, A^x, A^y, S^{ref}, S^x and S^y were constants to be determined. Using the known coordinates of the three pointed locations, we obtained three systems of linear equations:

$$\begin{cases} R/A/S(i_1,j_1) = R/A/S^{ref} + i_1 * R/A/S^x + j_1 * R/A/S^y \\ R/A/S(i_2,j_2) = R/A/S^{ref} + i_2 * R/A/S^x + j_2 * R/A/S^y \\ R/A/S(i_3,j_3) = R/A/S^{ref} + i_3 * R/A/S^x + j_3 * R/A/S^y \end{cases} \quad \text{Eq. 53}$$

where R/A/S was used to denote either the R, A or S coordinates. Assuming that the pointed location are not colinear, which was explicitly asked to the neurosurgeon, these three linear equation systems could be solved using the Cramer's rule (Cramer, 1750):

$$R/A/S^{ref} = \frac{\det\left(\begin{bmatrix} R/A/S(i_1,j_1) & i_1 & j_1 \\ R/A/S(i_2,j_2) & i_2 & j_2 \\ R/A/S(i_3,j_3) & i_3 & j_3 \end{bmatrix}\right)}{\det(B)} \quad \text{Eq. 54}$$

$$R/A/S^{x} = \frac{\det\left(\begin{bmatrix} 1 & R/A/S(i_1,j_1) & j_1 \\ 1 & R/A/S(i_2,j_2) & j_2 \\ 1 & R/A/S(i_3,j_3) & j_3 \end{bmatrix}\right)}{\det(B)} \quad \text{Eq. 55}$$

$$R/A/S^{y} = \frac{\det\left(\begin{bmatrix} 1 & i_1 & R/A/S(i_1,j_1) \\ 1 & i_2 & R/A/S(i_2,j_2) \\ 1 & i_3 & R/A/S(i_3,j_3) \end{bmatrix}\right)}{\det(B)} \quad \text{Eq. 56}$$

$$\text{with } B = \begin{bmatrix} 1 & i_1 & j_1 \\ 1 & i_2 & j_2 \\ 1 & i_3 & j_3 \end{bmatrix} \quad \text{Eq. 57}$$

The positions of all the electrodes of the grid on the reconstructed cortical surface could thus be estimated by using their coordinates in the grid space (**Fig. 54**-B and **Fig. 56**-A).

7. Implementation details

All the algorithms used to analyze and map the neural activity in real-time were developped in C++. Great care was given to optimization and parallelization of all computationally expensive algorithms. Indeed, in a typical experimental setting, we could expect to have about 250 neural channels recording at about 10Khz, which represents about 100Mbits of incoming data per second that has to be analyzed in real-time using computationally expensive technique such as the STFT. The STFT transform of the neural data was computed using the FFTSS library (Nukada, 2006) and parallelized using the OpenMP library (http://openmp.org) to take advantage of all the processor cores of the computer the software was running on. The choice for the FFTSS library was motivated by preliminary comparisons with other available libraries, which showed that it was particularly efficient. This work was performed by Philemon Roussel, an intern that I supervised during my thesis. We used the linear algebra Lapack++ (http://lapackpp.sourceforge.net/) library to compute the thin plate spline interpolation, and the Boost library (http://www.boost.org/) to compute the Student's distribution.

In order to avoid numerical instability and limit the accumulation of rounding errors, the iterative formula of the mean μ and standard deviation σ was used to compute mean and

standard deviation spectrums of the neural signal x, using an intermediate variable S (Chan et al., 1983):

$$\mu(t_0) = x(t_0)\ and\ \sigma(t_0) = S(t_0) = 0 \quad \text{Eq. 58}$$

$$\mu(t_{k+1}) = \mu(t_k) + \frac{x(t_{k+1}) - \mu(t_k)}{k+1} \quad \text{Eq. 59}$$

$$\begin{cases} S(t_{k+1}) = S(t_k) + \dfrac{x(t_{k+1}) - \mu(t_k)}{x(t_{k+1}) - \mu(t_{k+1})} \\ \sigma(t_{k+1}) = \dfrac{S_{k+1}}{k} \end{cases} \quad \text{Eq. 60}$$

III. Results

1. ClientMap: a neural activity mapping software dedicated to speech

The previously explained algorithms were all implemented in a dedicated software, called ClientMap. This software receives in real-time the neural activity and other signals – here the patient audio signal – that is streamed by our recording software NeuroPXI, and then performs the mapping of the speech-related activity, from the automatic speech detection to the final display of the activity.

Preliminary experiments demonstrated that thanks to the various optimization methods we used, we were able to analyze in real-time neural activity recorded from 256 electrodes at 10kHz, with STFT windows up to 4096 samples and 90% overlap, on an Intel Xeon ES5-2640 processor, which was necessary to perform the per-operative mapping in real-time.

The final software was organized into various pannels for configurating the analysis and visualizing its results either on pictures of the operative field or of the reconstructed cortical surface (**Fig. 60**).

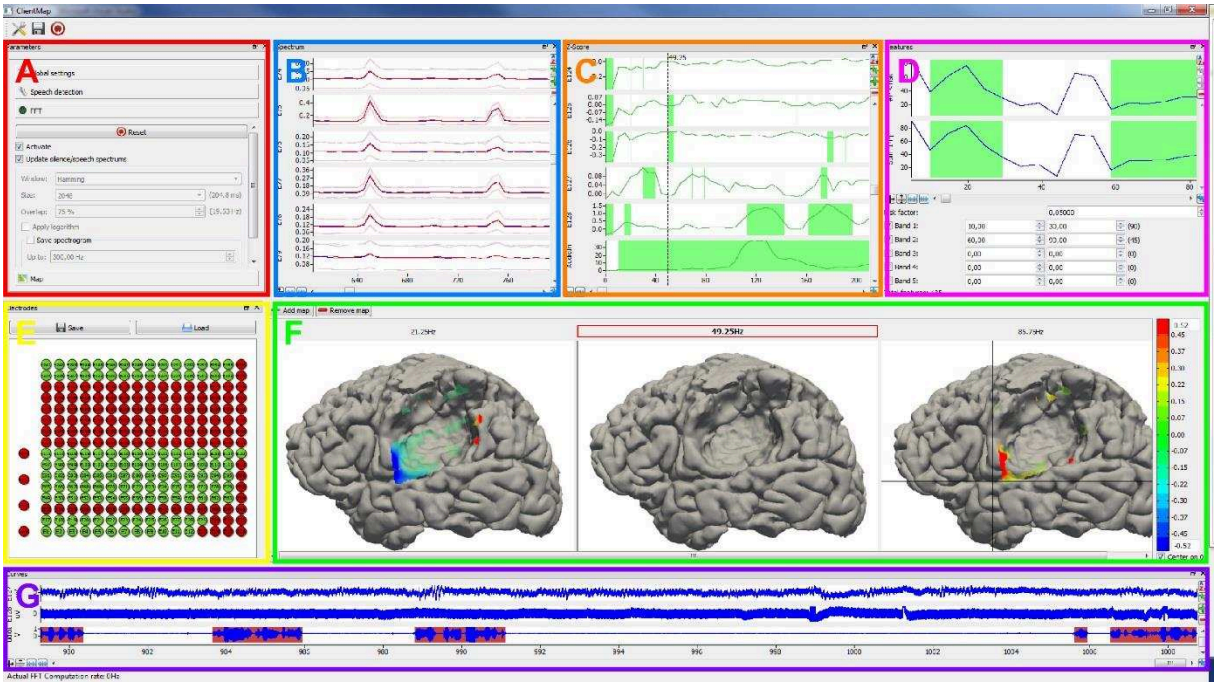


Fig. 60: Overview of the ClientMap software. A – Parameters panel. B – Spectrum panel. C – Score panel. D – Features panel. E – Electrodes panel. F – Maps panel. G – Raw data panel.

a. Parameters panel

The parameters panel is itself made of several sections (**Fig. 61**): (i) Global settings, (ii) Speech detection, (iii) FFT and (iv) Map.

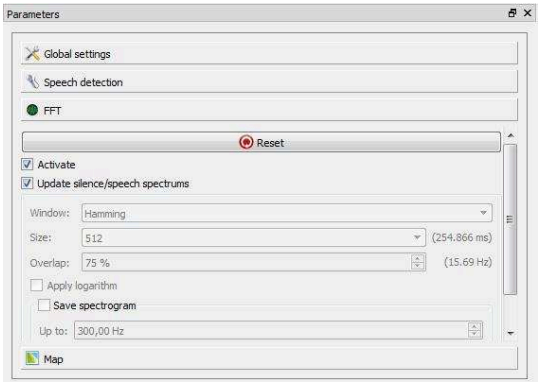


Fig. 61: Parameters panel of the ClientMap software. This panel contains several sections: Global settings, Speech detection, FFT and Map.

Global settings section. This section allows the experimentator to specify where to store data. Indeed, since the analysis is performed in real-time, results are changing over time. The software thus allows to save the current analysis state (parameters, maps, speech-silence-ratio, etc.) at any moment during the experiment.

Speech detection section. This section is used to specify the speech detection parameters (audio channel, threshold, inactivation period, etc.) and to enable or disable the automatic speech detection.

FFT section. This section is used to set the parameters of the short-time Fourier transform: window size, windowing function, overlaping between consecutive windows, etc. It allows as well to enable/disable the STFT computation and to reset the averaged spectrums. Indeed, if an artefact occurs during the analysis, for instance during a speech segment, it could invalidate the current analysis so that it must be resetted. While this was not done in the present work, automatic artefact rejection could be considered in future works to avoid the need for resetting the analysis.

Map section. This section allows to adjust the display settings for the maps (composition mode, opacity, gradient, display of the electrode positions, etc.). It allows as well to setup the coregistration of the electrodes on the anatomy.

b. Spectrum panel

The spectrum panel displays, for each recorded channel, the averaged signal sepctrum during silence and during speech, along with their respective standard deviation (**Fig. 62**). While this panel is not directly used to visualize the speech-related neural activity, it can be used for instance to quantify the electromagnetic interferences coming from the environment, essentially from power lines at 50Hz and its harmonics.

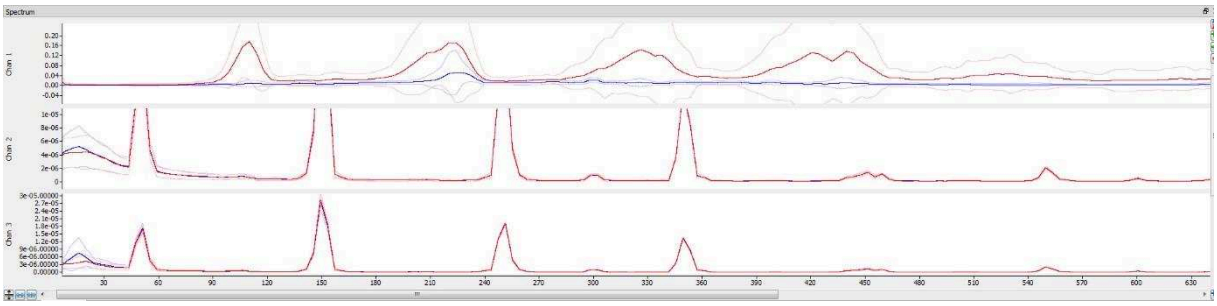


Fig. 62: Spectrum panel of the ClientMap software. This panel displays, for each electrode, the averaged spectrum of the neural activity during silence (blue curve) and speech (red curve) along with their standard deviation (semi-transparent blue and red curves).

c. Score panel

The score panel displays, for each recorded channel, the speech-silence-ratio at each frequency. Moreover, it also indicates when this ratio is significant by coloring the background of the displayed curved so that the user can quickly identify frequencies of interest for a specific channel (**Fig. 63**).

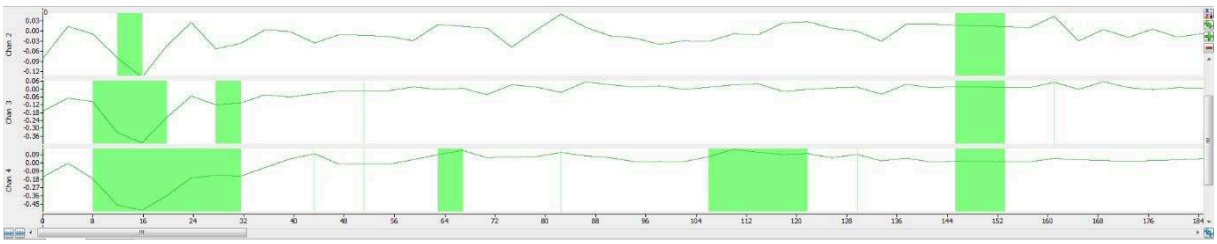


Fig. 63: Score panel of the ClientMap software. This panel displays, for each recorded channel, the speech-silence-ratio at each frequency (green curves). It as well indicates when this ratio is significant (green areas in the background of each curve).

d. Features panel

The features panel (**Fig. 64**) displays, for each frequency, the number of electrodes that exhibit significant speech-related activity. This information is useful to quickly identify frequency bands in which most electrodes exhibit significant speech-related activity. In this panel, the user can as well specify the risk factor α used for the statistical analysis in order to only map significant neural activity changes.

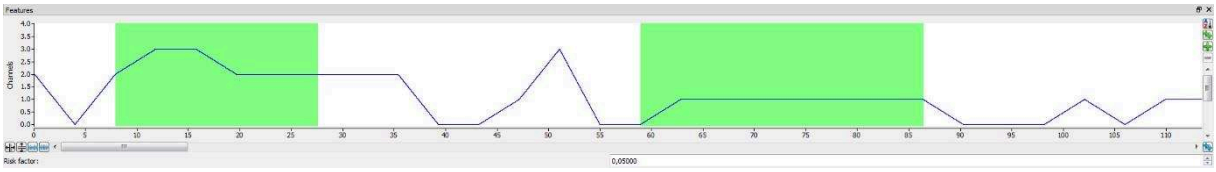


Fig. 64: Features panel of the ClientMap software. This panel displays, for each frequency, the number of electrodes that exhibit significant speech-related activity (blue curve). It is as well used to specify the uncorrected risk factor for the statistical analysis.

e. Electrode panel

The electrode panel allows to visualize the electrode grid geometry and channel identifiers, as well as to include or exclude any channel from the analysis. This is for instance used to exclude saturated or artefacted channels, and to exclude all other external channels that do not contain neural activity, for instance the patient audio channel (**Fig. 65**). The grid geometry is sent to the ClientMap software by the NeuroPXI software when starting the streaming of the data.

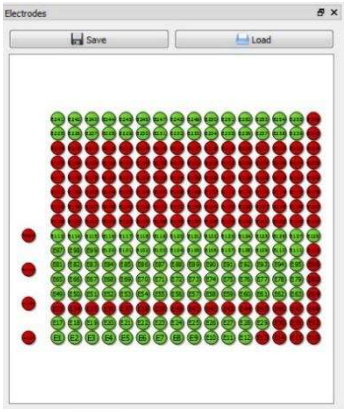


Fig. 65: Electrodes panel of the ClientMap software. This panel allows to visualize the electrodes (circles) grid geometry and channel identifiers (names in the circles), as well as to include (green circles) or exclude (red circles) any channel from the analysis.

f. Maps panel

The maps panel displays the speech-related brain activity at different frequencies (**Fig. 66**). Several maps can be added to visualize the activity for different frequencies at the same times. The maps are updated in real-time and are linked to the other displays so that the user can easily choose a frequency of interest for each map. Only significant values – according to the specified risk factor – are displayed.

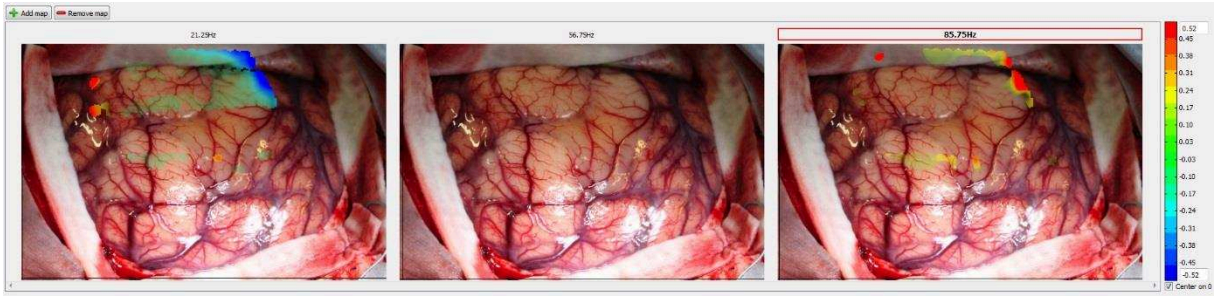


Fig. 66: Maps panel of the ClientMap software. This panel displays the speech-related brain activity at different frequencies (specified on top of each map). The color scale of all the maps can be adjusted using a dedicated interaction element (on the right).

g. Raw data panel

The raw data panel allows to visualize the incoming flow of data that is streamed from the NeuroPXI software (**Fig. 67**). This is useful for instance to detect saturated or artefacted channels and remove them from the analysis. On the patient audio channel, it as well displays which segments were automatically labeled as speech, in order to monitor the good functionning of the automatic speech detection.

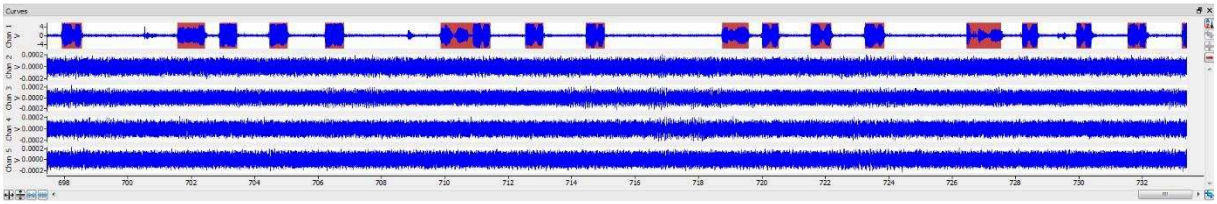


Fig. 67: Raw data panel of the ClientMap software. This panel allows to visualize the incoming flow of data (blue curves), as well as the audio channel segments that were automatically labeled as speech (red background).

2. Mapping of speech-related brain activity

a. First patient

i. Overt speech

The first patient had 4 recorded electrodes placed over the inferior prefrontal gyrus and its opercular part. For this patient the neural signal analysis was performed offline while simulating real-time condition by playing back the recorded data. **Fig. 68** shows an example of recorded neural signal.

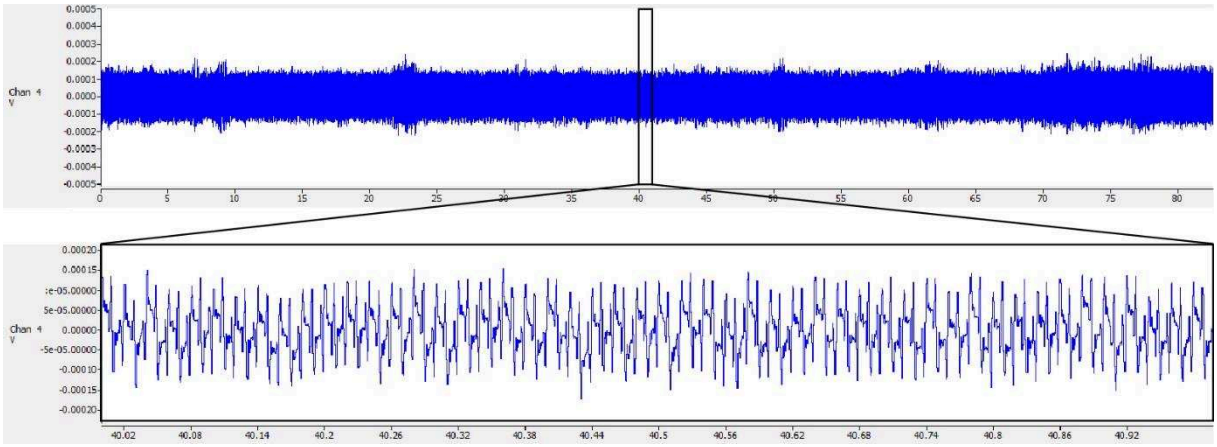


Fig. 68: Example of recorded neural signal for the first patient. There is a high noise level, especially due to environmental electromagnetic interferences at 50Hz (bottom row).

The analysis of the C value (i.e. the number of electrodes exhibiting significant speech-related activity, see "Extraction of speech-related brain activity" in the Methods) allowed to identify two principal frequency bands for which there was a power difference between speech and silence conditions for a significant amount of electrodes, corresponding to the beta band (from about 10Hz to 30Hz) and the low gamma band (from about 60Hz to 90Hz) (**Fig. 69**).

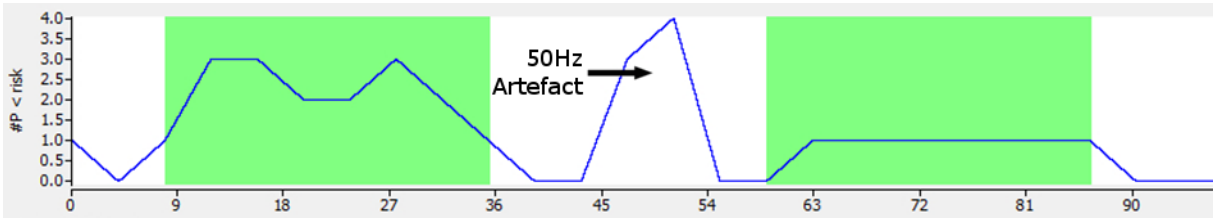


Fig. 69: Number of electrodes exhibiting significant speech-related activity for patient P1. For each electrode and frequency, *significant change in activity between speech and silence was assessed using a Welch's t*-test with Bonferroni risk correction. The curve shows, for each frequency (here from 0 to 90Hz), the number of electrodes which P-value was inferior to the *corrected risk factor (see "*Extraction of speech-related brain activity*" in the Methods).* The arrow shows a peak due to environmental electromagnetic noise (50Hz artefact).

Mapping the neural activity at these frequencies confirmed this result: a beta desynchronization was observed on two electrodes over the opercular part of the inferior prefrontal gyrus during speech production (**Fig. 70**-Left), as well as a significant increase of gamma activity in one of these two electrodes (**Fig. 70**-Right).

Fig. 70: Mapping of the speech-related activity for patient P1. Left – Beta desynchronization (here mapped at 16Hz) in the inferior precentral sulcus and anterior subcentral sulcus during speech production (blue area). Note that the red areas are not relevant here since they are the results of an extrapolation outside fo the electrodes grid. Right - Increase of gamma activity (here mapped at 76Hz) in the inferior precentral sulcus and anterior subcentral sulcus (red area).

This result could be quickly observed after the first pronounced items and was maintained until the end of the experiment. The time-frequency representation of the neural signal for the particular that exhibited an increase in the gamma band (**Fig. 70**) was also averaged across all trials (**Fig. 71**).

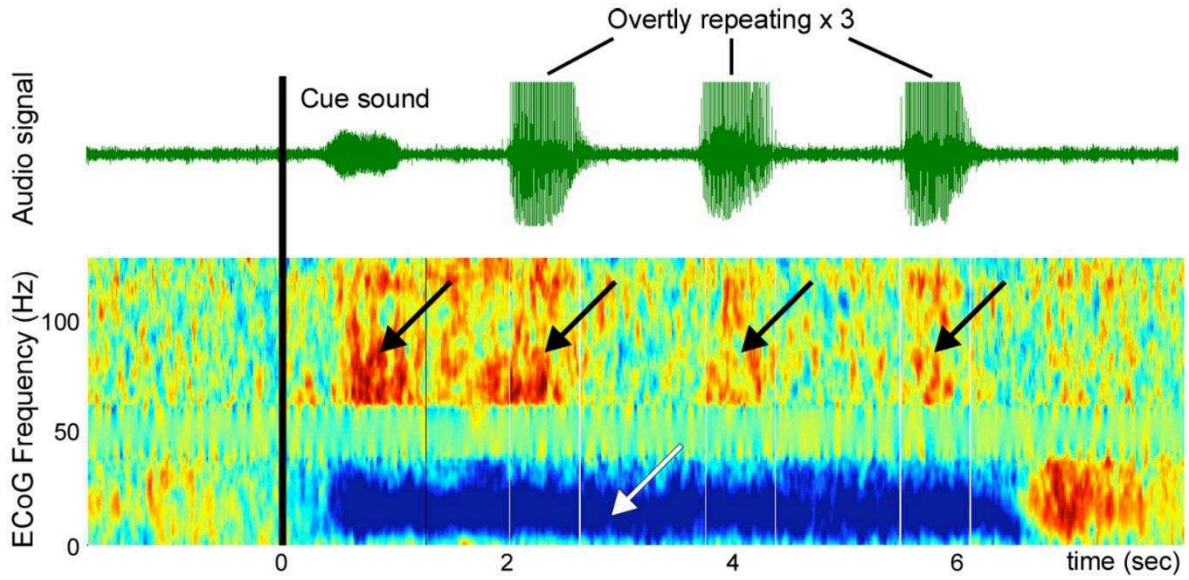


Fig. 71: Average time-frequency representation of the neural data during overt speech. Top – Sample sound recorded by a microphone placed next to the awake patient. Bottom – Time-frequency representation of the ECoG signal showing clear beta desynchronization (blue, white arrow) and gamma-band responses to the cue and for each pronounced sound (red, black arrows). ECoG data was recorded at 2 kHz and the time-frequency representation was computed using short-time Fourier transform using a Hamming function on 512 samples sliding windows with 95% overlap. The time-frequency representation was then normalized by the 1-sec pre-stimulus period and averaged over 83 trials aligned on the beginning of the cue signal.

The time-frequency representation of the neural signal exhibits a clear desynchronization in the beta band (10-30Hz) that starts at the beginning of the cue sound, is sustained during the whole speech production, and is followed by a rebound at the end of the task (**Fig. 71**, white arrow). An increase of power was also observed in the low and high gamma bands (60-130Hz) during the cue sound, and also during each pronounced occurrence (**Fig. 71**, black arrows). As opposed to the desynchronization in the beta-band, the gamma-band increase seems to not be sustained between each speech production.

ii. Covert speech

The first patient also produced covert speech during the experiment. He was first asked to listen to the same speech item played three times at a fixed pace and then to continue to imagine repeating it at the same pace. The end of each imagination period was notified by the patient who pronounced the word "ok" aloud. We thus also averaged the time-frequency representation of the neural signal from the previously identified speech-specific electrode occurences (**Fig. 72**).

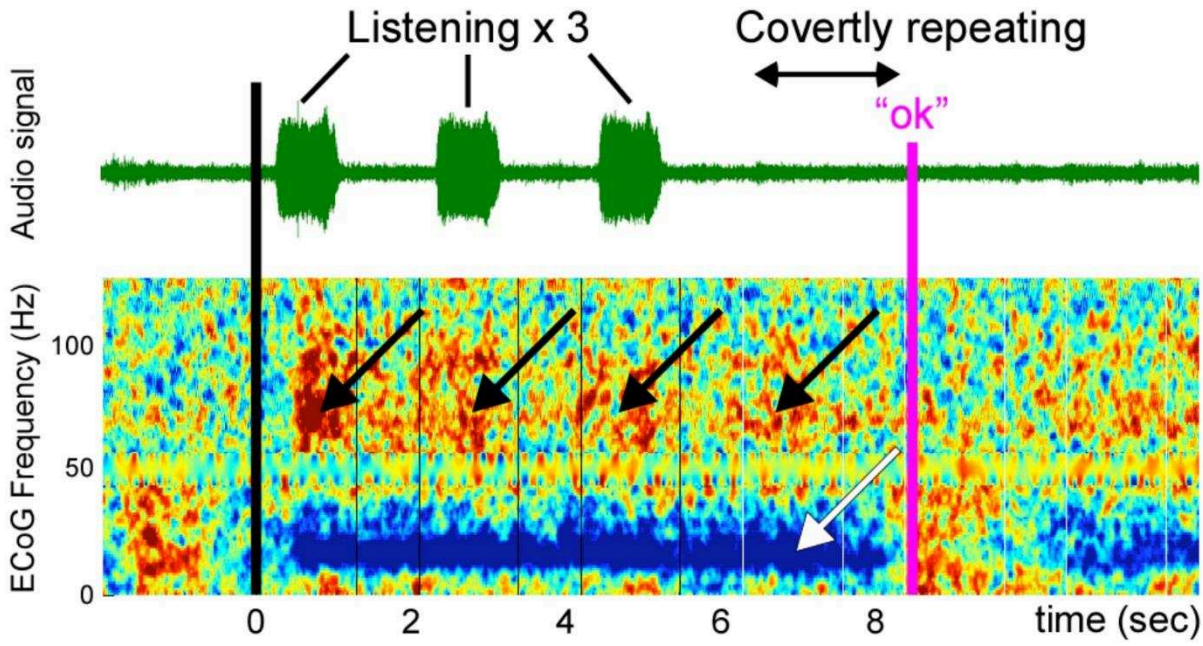


Fig. 72: Average time-frequency representation of the neural data during covert speech. The modulation of beta (white arrow) and high-gamma band activity (black arrows) over the speech motor cortex during speech listening is prolonged during the period the subject is asked to imagine repeating what he has heard. Top – Sound recorded by the microphone positioned next to the awake patient. Bottom – Time-frequency representation of the ECoG signal averaged over 24 trials on the same electrode and using the same methods as in **Fig. 71**. The vertical pink line shows the mean position of the end of the imagination period as notified by the patient by saying "ok" aloud.

Only 24 items were averaged for the covert speech representation, while 84 items were used for overt speech. However, similarities between overt and covert speech could still be observed. As for overt speech, there is a clear beta desynchronization (10-30Hz) during the listening of the speech items that is sustained during the speech imagination period (**Fig. 72**, white arrow), and is also followed by a rebound (**Fig. 72**, increase after the vertical pink line). There is also an increase in the low and high gamma bands (60-130Hz) during the listening period that seemed to be prolonged during speech imagination (**Fig. 72**, black arrows).

b. Second patient

The second patient had 256 recorded electrodes, mostly placed over its tumor, which was located in the left central silcus, with some electrodes covering surrounding areas, notably on the lower part of the motor and sensory cortices (**Fig. 56**-A). However, technical difficulties (principally due to the rigidity of the grid because of the high number of electrodes) drastically reduced the number of electrodes that were correctly recorded: 125 electrodes out of 256 had to be removed fom the analysis, either because they were saturated or because there was no signal at all, which resulted in having almost no electrodes over the sensory motor cortex. In total, about 15 minutes of continuous neural and audio data were considered for an offline analysis after artefact removal.

Among the correctly recorded electrodes, most were directly located over the tumor and did not show any speech-specific neural activity, as could be expected. Similarly to patient P1, the analysis of the C value (i.e. the number of electrodes exhibiting significant speech-related activity, see "Extraction of speech-related brain activity" in the Methods) allowed to identify two principal frequency bands for which there was a power difference between speech and

silence conditions for a significant amount of electrodes, corresponding to the beta band (from about 10Hz to 30Hz) and the low gamma band (from about 60Hz to 90Hz) (**Fig. 73**).

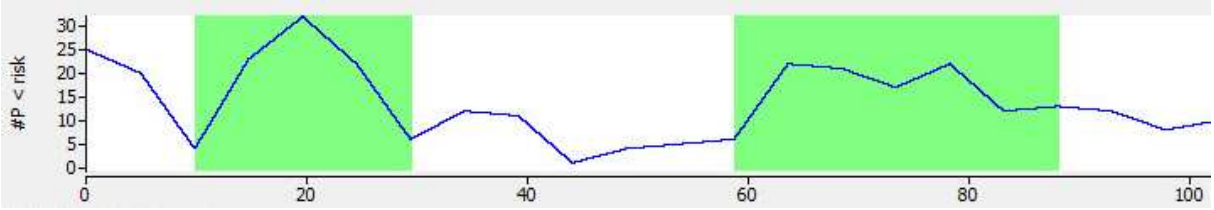


Fig. 73: Number of electrodes exhibiting significant speech-related activity for patient P2. For each electrode and frequency, *significant change in activity between speech and silence was assessed using a Welch's t*-test with Bonferroni risk correction. The curve shows, for each frequency (here from 0 to 100Hz), the number of electrodes which P-value was inferior to the corrected risk factor *(see "*Extraction of speech-related brain activity*" in the Methods)*.

This was confirmed by mapping the neural activity at different frequencies. In particular, the mapping showed that there was a large beta desynchronization in the inferior precentral gyrus (Brodmann area 4) and anterior subcentral gyrus (Brodmann area 43) during speech production (**Fig. 74**-A), as well as a smaller but overlapping increase of gamma activity in the same area (**Fig. 74**-B). Note how both these activities seem to expand to more frontal parts of the brain that could not be covered by the ECoG grid in this experiment.

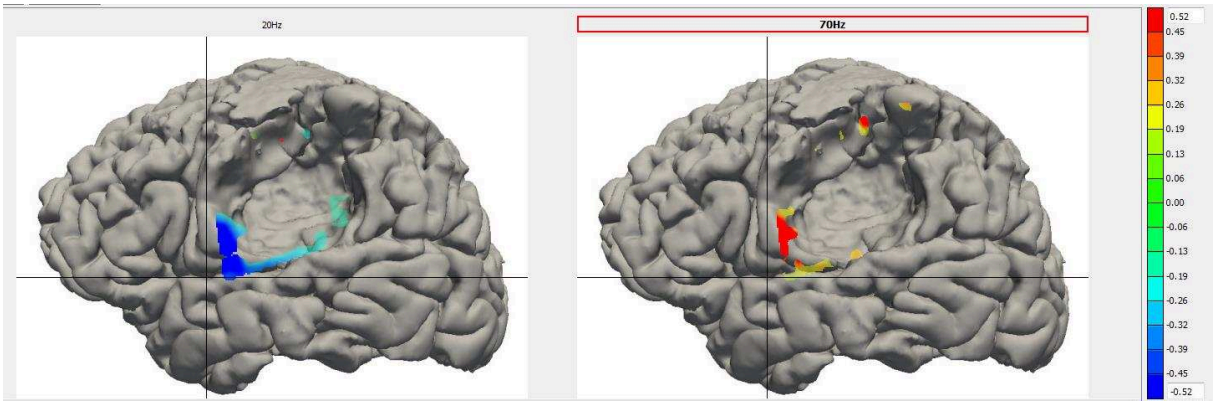


Fig. 74: Mapping of the speech-related activity for patient P2. Left – Beta desynchronization (here mapped at 20Hz) in the inferior precentral sulcus and anterior subcentral sulcus during speech production (blue area). Right - Increase of gamma activity (here mapped at 70Hz) in the inferior precentral sulcus and anterior subcentral sulcus (red area).

As for P1, these results could be observed after a few pronounced items and were stable all along the whole experiment.

IV. Conclusion

In this chapter, we first presented a method to localize and visualize speech-related brain activity in real-time, during awake brain surgery. This method was then implemented in a dedicated software, called ClientMap. This software was tested on two subjects undergoing brain surgery for a tumor removal in order to localize and visualize speech-related brain areas, especially in the motor cortex.

Results were consistent for both patients, and showed that a beta desynchronisation as well as an overlapping increase in gamma activity occur over the speech motor cortex during speech

production with respect to rest activity. More precisely, the beta desynchronization was observed on two electrodes located over the opercular part of the inferior prefrontal gyrus and the gamma activity on one of these two electrodes for the first patient (**Fig. 70** and **Fig. 71**). For the second patient, the beta desynchronization was observed in the inferior precentral gyrus and anterior subcentral gyrus and the increased gamma activity, while less spread out, in the same area (**Fig. 74**). Interestingly, this similar activity between both patients could be observed despite crucial differences in the two experimental protocols. Indeed, the first patient was asked to produce only isolated vowels or vowel-consonant-vowel sequences, while the second patient produced full speech sentences in a continuous paradigm. Moreover, the speech items were orally presented to the first patients at a predefined pace through speakers, while they were visually presented on a screen for the second patient, that could read them at the pace he desired. This suggests that different speech production tasks still elicit similar neural activity in the speech motor cortex. This should however be confirmed by future experiments. In particular, analysis of the neural activity could be done at the phone level. Using automatic speech recognition methods, it could be possible to directly segment the incoming speech signal into phones or places of articulation. This could be used to more precisely map the speech neural activity for specific speech sounds or specific speech articulators.

Moreover, neural activity similar to that of overt speech was observed during covert speech. Indeed, for the first patient, a beta desynchronization was also observed at the same location as during overt speech and was sustained during the whole imagination period, as well as a slight increase in the low and high-gamma bands (**Fig. 72**). This suggests that overt and covert speech at least share a partially common neural representation in the speech motor cortex. Future experiments should also confirm this hypothesis by using denser recordings. Indeed, the relatively large size and the limited number of the electrodes used in the first patient did not allow to precisely identify the similarities between overt and covert speech neural activities, in particular for the low and high gamma bands.

However, the mapping approach we proposed here can be taken further in future works by improving several points. First, the automatic detection of speech has to be improved to avoid undetecting low energy sounds at the beginning or end of a speech utterance. This can be done in many ways, such by using a voice-activity detector based on the statistical modelling of silence versus speech content, for instance using hidden markov models (HMMs) or Gaussian mixture models (GMMs). Secondly, the mapping workflow might be improved by automatizing or semi-automizing the coregistration of the electrodes on the anatomy. While the approach we propose here is rather fast since it only requires the localization of a dozen electrodes and anatomical landmarks, the experimental time available during awake surgeries is very limited (about 30 minutes), and could greatly benefit from automatizing this process. The localization of the electrodes on the picture with the ECoG grid visible could be semi-automatically achieved by letting the user identify three non-colinear electrodes on the pictures, which would then be used to statistically model the electrode pixel distribution to detect all the other electrodes, and then combined with the known geometry of the ECoG grid to assign to each electrode its correct identifier – for instance its channel number. Pairs of correspondence points between the image with and without the grid visible could be as well automatically identified. Algorithms from the computer vision field could be used to identify pairs of matching points, for instance using scale- and rotation-invariant features such as SURF (Bay et al., 2006) or SIFT (Lowe, 1999) features, and mathematical camera models taking into account lens effect and

perspective projection could be used to improve the quality of the matches. Such methods were not developped during this thesis but should be considered in the future. Additionally, the localization of the electrodes on the reconstructed brain surface might also be improved. In particular, we did not take into account the curvature of the electrode grid and considered that all the electrodes were contained in the same 3d plane. Future work could take into account the reconstructed cortical surface geometry in order to better estimate the locations of the electrodes by constraining them to be on the surface of the brain.

Altogether, these results support the fact that ECoG can be used during brain surgery to precisely localize and map speech-related brain areas directly on the exposed cortex. For both patients, speech-related neural activity could be mapped within the first minutes of the experiment and was stable during the whole experiment. Such mapping could thus be used to identify functional areas that should be preserved directly during the surgery and prior to the resection. Moreover, this mapping can also help identify best candidate areas for the positioning of micro-electrode arrays for speech decoding.

Chapter 7: Speech decoding from neural activity

I. Introduction

Neural activity from the speech-related brain areas can be used to decode speech in order to restore communication in aphasic patients, such as locked-in patients. Most studies aiming at restoring speech from neural activity considered the decoding of acoustic features (Guenther et al., 2009; Martin et al., 2014) or phonetic units (Kellis et al., 2010; Brumberg et al., 2011; Pei et al., 2011a; Tankus et al., 2012; Mugler et al., 2014; Herff et al., 2015), while few studies considered articulatory features (Pasley and Knight, 2013; Lotte et al., 2015).

Therefore, one goal of my thesis was to start investigating how articulatory trajectories could be decoded from neural activity. In this chapter, we thus considered the decoding of electromagnetic articulography (EMA) data, as presented in **Chapter 3**, from neural activity recorded in the speech motor cortex (see **Chapter 6**). While EMA data provides enough information to synthesize intelligible speech (see **Chapter 4 and 5**), it does not provide any information about speech intention, i.e. EMA data alone cannot be used to decide when to synthesize or not synthesize speech according to the patient intent. In this chapter we thus as well considered speech intention detection from neural data, i.e. the prediction of the intention or not to speak from neural data. Moreover, the addition of voicing information can help improve the synthesis intelligibility, especially for consonants (see **Chapter 4**). Thus, we as well considered the prediction of a binary voicing activity, i.e. predicting if the larynx is vibrating to produce voiced speech (such as vowels), or on the contrary if it produces unvoiced speech (such as the consonant /s/).

The results reported in this chapter are preliminary and still under development.

II. Methods

1. Decoding of speech intention

As a first step toward speech decoding, we considered the decoding of speech and non-speech states, i.e. predicting from the neural activity alone whether the patient is producing or intent to produce speech or not. Speech intention detection has been previously investigated with success (Kanas et al., 2014), but only using data recorded during overt speech, which is why it was referred to as voice activity detection in these studies. Since we consider here all kinds of speech production (overt, silent and covert), we prefer the term speech intention detection. This kind of “brain switch” would be particularly usefull in a future BCI for speech rehabilitation in order to detect whether the patient intends to speak or not.

a. Subjects and experimental design

Subjects and experimental design are identical to **Chapter 6**. Please note that only the first patient performed covert speech. For both patients, speech and non-speech segments were manually labeled.

b. Features extraction

i. First patient

For the first patient, only the data from one electrode that was shown to be speech-specific was kept (**Fig. 75**, see also **Chapter 6**).

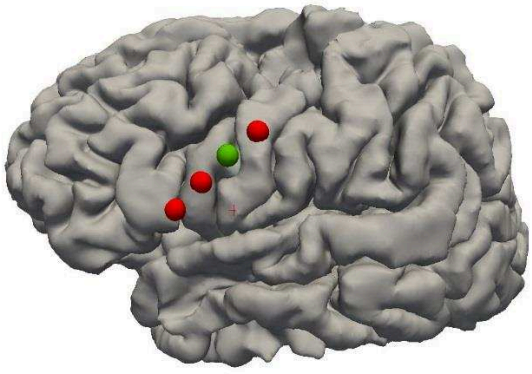


Fig. 75: Electrodes showing speech-specific activity used for the decoding for the first patient. Only the electrodes that exhibited speech-specific activity were considered for the decoding. One electrode localized next to the speech motor cortex was selected (in green).

The neural data was first segmented into speech and non-speech segments. Non-speech segments were randomly chosen to obtain the same number of speech and non-speech frames. The neural data was then continuously segmented into 128ms windows (256 samples at 2kHz) shifted by 10ms (i.e. sampled at 100Hz). Each signal window was then windowed by a Hamming window and processed using short-term Fourier analysis (STFT) to extract the signal power in frequency bins of approximately 2Hz (513 bins in total). The data during rest periods (1 second period before each stimuli) was used to normalize with a z-score the signal power in each frequency bin. Only the normalized power for frequencies in the beta and gamma bands was kept (from 10Hz to 90Hz). At the end of this step, the neural signal was thus represented by 41 different features, sampled at 100Hz.

The same process was applied both to the overt and to the covert data.

ii. Second patient

For the second patient, electrodes were first re-referenced to a common average (i.e. for each sample and channel, substracing the mean of all the valid channels) to remove slow variations and common artefacts, such as power-line interferences, saturated electrodes were excluded. Then, electrodes which neural activity was speech-related were localized (see **Chapter 6, Fig. 74**), and only electrodes with significant speech-specific activity (with a Bonferroni corrected p-value inferior to 0.05) were kept. This allowed to reduce the subset of

electrodes used for the decoding to a total of 20 channels localized over the speech motor cortex (**Fig. 76**, see also **Chapter 6**).

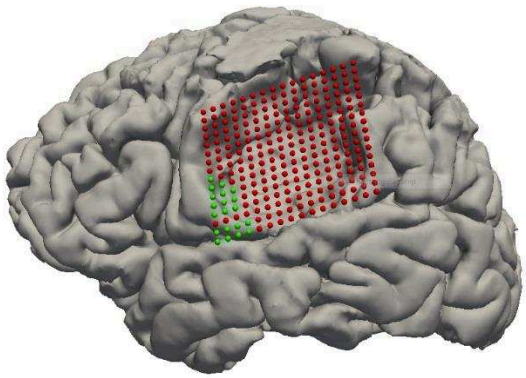


Fig. 76: Electrodes showing speech-specific activity used for the decoding for the second patient. Only the electrodes that exhibited speech-specific activity were considered for the decoding. Twenty electrodes localized over the speech motor cortex were selected (green dots). Some electrodes in this area were excluded because of their high noise level.

The neural data was then continuously segmented into 205ms windows (2048 samples at 10kHz) shifted by 10ms (i.e. sampled at 100Hz). Each signal window was then windowed by a Hamming window and processed using short-term Fourier analysis (STFT) to extract the signal power in frequency bins of approximately 5Hz (1024 bins between 0Hz and 5kHz). The data during rest periods was used to z-score the signal power in each frequency bin. Then, the mean power over 10Hz-wide frequency bands from beta to gamma activity was computed (10-20Hz, 20-30Hz, 60-70Hz, 70-80Hz and 80-90Hz), resulting in 5 features per electrode every 10ms.

The slow potential of each electrode was then added to this set of features by band-pass filtering the raw signal between 0.5 and 5Hz (Kornhuber and Deecke, 2016). At the end of this step, the neural signal was thus represented by 120 different features (20 electrodes and 6 features per electrode), sampled at 100Hz.

c. Classification method

The classification of speech and non-speech segments was performed using support vector machines (SVM) with the radial basis function as kernel function (Chang et al., 2010). Originally, SVM is a supervised machine learning method that can be used to perform binary – i.e. in two classes – linear classification. SVM works by considering each data point as a point in a N-dimensional space, and by finding the hyperplane that best separates labeled data points from a training set (i.e. the class they belong to is known). Such hyperplane is chosen so that it lies withing the larger gap that separates both classes, i.e. so that each class points are on the opposite side of the plane than the other class points, and that the distance between the plane and the closest point of each class – called the margin – is the greatest. This separation boundary allow to classify new data points according to their relative position to the hyperplane. Choosing a plane with a high margin allows to reduce chances of misclassification (i.e. of attributing the wrong label to a new data point).

However, real-life problems are generally non-linear, i.e. two classes can generally not be separated by an hyperplane. Non-linear SVM can be performed by applying a function φ to the input data that transforms the input features space into a higher dimensionality space in which

the classes could be more easily separated using linear SVM. However, mapping data to very-high dimensionality spaces can be memory and computationally expensive. The "kernel trick" allows to avoid this explicit mapping by relying on the fact that only the pairwise dot-product $\varphi(x) \cdot \varphi(y)$ of the data points x and y mapped to the high dimensionality space is needed. Indeed, some functions k – called kernels – allow to implicitly compute the cross-product $\varphi(x) \cdot \varphi(y)$ in the original space so that:

$$k(x, y) = \varphi(x) \cdot \varphi(y) \quad \text{Eq. 61}$$

Here we used the commonly used radial basis function (RBF) as kernel function k:

$$k(x, y) = \exp(-\frac{\|x-y\|^2}{2\gamma}) \text{ with } \sigma > 0 \quad \text{Eq. 62}$$

The value of the γ parameter was chosen by testing different values (from 0.1 to 1000) and keeping the value that gave best results. We used the Matlab SVM implementation from the LibSVM library (https://www.csie.ntu.edu.tw/~cjlin/libsvm).

For the first patient, the decoding model that was only trained on the overt speech data was also applied to the covert speech data without any further modification, in order to assess if it could also predict speech intention.

d. Evaluation of the speech state decoding

The decoding was performed in a 5-fold cross-validation process, so that the original data was randomly divided into 5 chunks, 4 of which were used for training, while the remaining one was used for testing, which was repeated 5 times in order to test all partitions. The chance level was estimated by shuffling the data labels prior to training and test, which was repeated 20 times for each fold, and thus 100 times for each value of the γ parameter.

The decoding quality was assessed by computing the numbers of true positive TP (the number of speech frames correctly decoded as speech), true negative TN (the number of silence frames correctly decoded as silence), false positive FP (the number of silence frames decoded as speech) and false negative FN (the number of speech frames decoded as silence). This allowed to compute three different measures of the decoding quality: the sensitivity, the specificity and the accuracy:

$$Sensitivity = \frac{TP}{TP + FN} \quad \text{Eq. 63}$$

$$Specificity = \frac{TN}{TN + FP} \quad \text{Eq. 64}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{Eq. 65}$$

The accuracy reflects the global performance of the decoding, while the sensitivity reflects the correct detection of speech segments, and the specificity that of the silence segments. A value of 1 for each of these measures would correspond to a perfect decoding, while lower values denote errors in the decoding.

2. Decoding of the voicing activity

A second step was to decode the voicing activity into a binary state: voiced or unvoiced sounds. Additionally, decoding the voicing activity could help improve synthesis intelligibility (see **Chapter 4**).

a. Subjects, experimental design and features extraction

For the voicing activity decoding, only the second patient's data was considered.

First, the pitch was extracted from the original patient's audio using SPTK (see **Chapter 4**). The pitch is a time-varying value describing the voicing property of speech. A null pitch represents unvoiced periods, while a non-null pitch represents the period of the excitation signal produced when the larynx is vibrating. Here we considered a simpler binary feature, by grouping pitch values into two class: null values were grouped in the unvoiced class, and non-null values in the voiced class.

The neural data was processed identically as for the decoding of speech and non-speech states, at the exception that only speech segments were considered. Moreover, to take into account the dynamic properties of speech, we concatenated each neural data frame with its 9 preceding frames (100-ms time window context). Finally, voiced and unvoiced segments were randomly chosen to ensure an equal number of voiced and unvoiced frames.

b. Classification method and results evaluation

The decoding of the voicing was performed and evaluated using identical methods as for decoding of speech and non-speech states: the classification was done using SVM with a radial basis kernel function of different γ values, and the classification quality was assessed by measuring the accuracy, specificity and sensitivity.

3. Decoding of articulatory features

A third step was to test the extent to which all articulatory trajectories could be decoded from ECoG data. Indeed, once speech intention is detected, neural activity could be decoded into articulatory parameters that control an articulatory-based speech synthesizer to produce speech, such as the one developped in **Part 3**.

a. Subjects and experimental design

For the decoding of articulatory features, only the second patient's data was considered. Indeed, the number of recorded electrodes in the first patient that exhibited speech-specific neural activity – two electrodes – was too limited to expect decoding articulatory movements from neural data.

b. Neural features pre-selection and extraction

The neural data was first pre-processed as for the decoding of speech and non-speech states, and was thus represented by 120 different features, sampled at 100Hz.

Some components of brain activity precede the execution of a motor task by several hundreds of milliseconds (Kornhuber and Deecke, 2016). We thus considered different delays in the neural data by temporally shifting the data with 9 different equally spaced delays from 200 milliseconds before to 200 milliseconds after the original data time (-200ms, -150ms, -100ms, -50ms, 0ms, +50ms, +100ms, +150ms and +200ms). Moreover, in order to take into account the dynamics of neural activity, we concatenate each neural data frame with the N previous ones. Six different values of the context size N were tested: 0, 1, 2, 4, 8 and 16.

However, such high number of features can generally lead to overfitting issues[1], especially when the available data is limited. Thus, we investigated how dimensionality reduction could help improve the decoding accuracy by using principal component analysis (PCA), with varying number of components: 5, 10, 15, 20, 25, 30, 40, 50, 75 and 100 ; and without PCA.

Thus in total 594 different parametrizations (9 delays, 6 context sizes and 11 PCA configurations) were tested for the neural data. The silences were excluded so that only the speech data was considered for the decoding.

c. Estimation of articulatory features

While video recordings and ultrasonography could be used together to record articulatory data simultaneously with brain activity (Bouchard et al., 2016), here no articulatory data was recorded during the experiment. However, the patient pronounced sentences that were present in our BY2014 articulatory-acoustic corpus (see **Chapter 3**). Therefore, for these particular sentences, the articulatory inputs from the speech synthesizer were known (i.e. the articulatory trajectories recorded from the reference (BY2014) speaker, used to build the articulatory-based synthesizer).

However, since the sentences were only visually presented to the patient, important differences in manner of articulation might occur between the patient and the reference audio thus preventing a good alignment of the neural activity with the corresponding articulatory movements. To compensate for this, dynamic time warping (DTW) was used to automatically align the reference audio on the patient audio, as shown on **Fig. 77** (Sakoe and Chiba, 1978). DTW works by applying a non-uniform time warping to the signal to be aligned in order to obtain the minimum distance between the warped signal and a reference signal. Here we used the mel-cepstral distortion (see "Evaluation of the speech synthesis intelligibility" in **Chapter 4**) as the local distance.

[1] Note that this is generally not an issue when performing classification with SVMs and was thus not critical for the decoding of speech and non-speech state, as well as the decoding of the voicing activity. However the decoding model used here for the prediction of articulatory trajectories could be subject to overfitting issues.

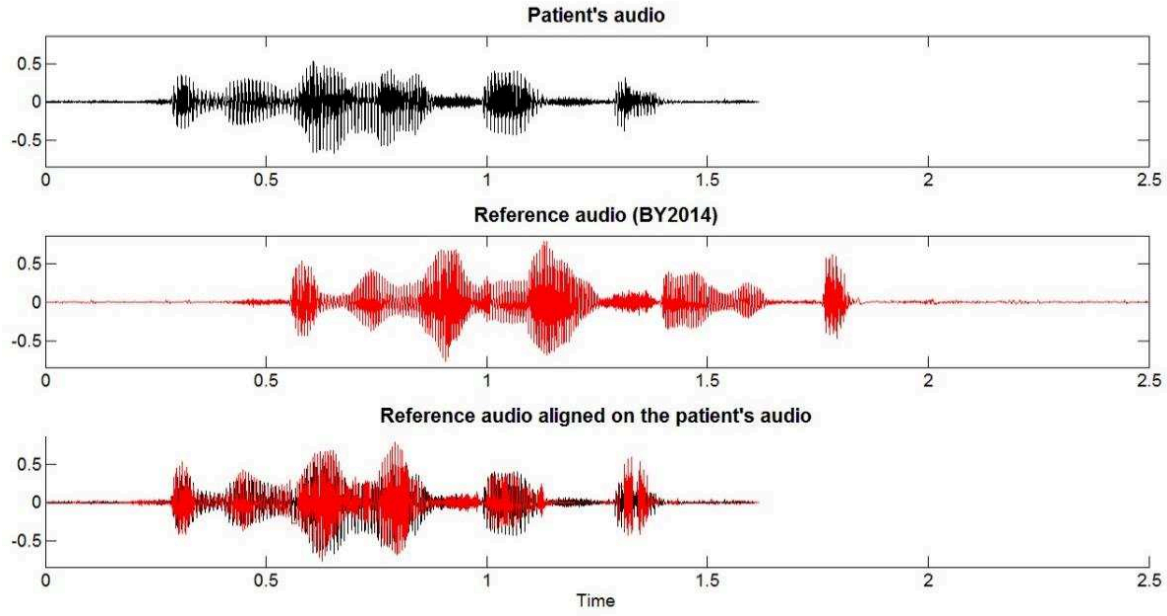


Fig. 77: Alignment of the reference audio from the BY2014 corpus on the patient's audio. Top row – the patient audio recorded during surgery. Middle row – The corresponding reference audio from the BY2014 corpus. Bottom row – the reference audio (red) is aligned on the *patient's audio (black) after applying DTW.*

Quality of the DTW was manually assessed by comparing the patient audio with the aligned reference audio, and all misaligned sentences were excluded, resulting in a total of 118 sentences left for the decoding.

The time-warping estimated using DTW was then applied to the reference articulatory data in order to align it on the neural signal. This allowed to obtain, for each sentence pronounced by the patient, an estimation of the corresponding articulatory data.

d. Neural-to-articulatory mapping

For this preliminary study, linear models were used to estimate the articulatory trajectories from the neural data. While more complex models could be used, such as GMMs or DNNs, this choice was motivated in order to limit overfiting effects due to the small size of the dataset (100 sentences) and the high features dimensionality of the inputs (up to several thousands features).

Thus, mapping the neural activity $\underline{\mathbf{X}}$ to articulatory trajectories $\underline{\mathbf{Y}}$ consisted in finding the matrix $\underline{\mathbf{A}}$ solving the equation:

$$\underline{\boldsymbol{Y}} = \underline{\tilde{\boldsymbol{X}}}.\underline{\boldsymbol{A}} \text{ with } \underline{\tilde{\boldsymbol{X}}} = [\underline{\boldsymbol{X}} \quad \underline{\mathbf{1}}] \qquad \text{Eq. 66}$$

If $\underline{\tilde{\boldsymbol{X}}}$ is inversible, the solution is given by:

$$\underline{\boldsymbol{A}} = \underline{\tilde{\boldsymbol{X}}}^{-1}.\underline{\boldsymbol{Y}} \qquad \text{Eq. 67}$$

However, $\underline{\tilde{\boldsymbol{X}}}$ is not inversible in most cases so that A is chosen by trying to minimize:

$$\left\|\underline{\boldsymbol{Y}} - \underline{\tilde{\boldsymbol{X}}}.\underline{\boldsymbol{A}}\right\| \qquad \text{Eq. 68}$$

Which is an overdetermined least squares problem having more than a unique solution. Here we estimated $\underline{\mathbf{A}}$ using the Moore-Penrose pseudo-inverse pinv:

$$\underline{\boldsymbol{A}} = \boldsymbol{pinv}(\underline{\tilde{\boldsymbol{X}}}).\underline{\boldsymbol{Y}} \quad \text{Eq. 69}$$

e. Evaluation of the decoding

For each set of parameters (number of PCA components, delay and context size), the mapping was performed in a 5-fold cross-validation process, so that the original data was randomly divided into 5 chunks, 4 of which were used for training, while the remaining one was used for testing, which was repeated 5 times in order to test all partitions.

Estimation of the chance levels was done by randomly shuffling the training data before estimating the linear model, then testing the model on the test data, randomly shuffled as well. This process was repeated 10 times for each cross-validation fold, and thus 50 times for each set of parameters (among the 594 different ones).

The decoding accuracy was then assessed by measuring the mean correlation between the predicted trajectories and the target trajectories.

4. Decoding of acoustic features

In order to better assess the hypothesis that decoding articulatory trajectories might be more revelant than decoding acoustic parameters when recording neural activity from the speech motor cortex, we also directly decoded acoustic parameters from the neural data.

a. Neural-to-acoustic mapping

The same procedure than for articulatory trajectories was applied, at the exception that the predicted signals were the mel-cepstrum coefficients directly extracted from the patient's audio signal (using SPTK) instead of the estimated articulatory trajectories. In order to be consistent with the BY2014 articulatory-acoustic corpus, the patient audio was resampled at 22kHz, and 25 mel coefficients were extracted and sampled at 100Hz (see **Chapter 3**).

As for the decoding of articulatory features, silences were removed so that only speech segments were considered for the decoding.

b. Comparison with the neural-to-articulatory mapping

Signals of different nature – such as the predicted articulatory and acoustic features – cannot be directly compared. Thus, the articulatory features predicted using the neural-to-articulatory mapping were mapped to acoustic features using the articulatory-to-acoustic mapping of the articulatory-based synthesis (see **Chapter 4**). This allowed to obtain a similar acoustic representation for the neural-to-articulatory and the neural-to-acoustic mappings allowing further comparison.

However, in the case of the neural-to-acoustic mappings, the acoustic features represent the audio signal of the patient, while in the case of the articulatory-to-acoustic mapping, the acoustic features represent the audio signal of the BY2014 subject. Thus, we used the time-

warping of the DTW alignments (see "Estimation of articulatory features") to align as well the BY2014 acoustic features on the neural data. The acoustic features predicted using the neural-to-articulatory cascaded with the articulatory-to-acoustic mappings (called "neural-to-articulatory-to-acoustic mapping" in the following) were therefore compared with these BY2014 acoustic features considered as the ground truth, while the acoustic features predicted using the neural-to-acoustic mapping were compared to the patient's acoustic features.

Since many different representations of the neural data were used (by varying the delay, context size and number of principal components), this comparison was performed by only considering the representation that led to the best decoding, i.e. the delay, context size and principal components for the neural-to-articulatory mapping was chosen as the combination that led to the highest correlation for each one of the 14 articulatory trajectories, while for the neural-to-acoustic mapping it was the combination that led to the highest correlation for each one of the 25 acoustic features.

III. Results

1. Decoding of speech intention

As a first step toward a speech BCI, we considered the decoding of speech and non-speech states for both patients.

a. First patient

The first patient produced both overt and covert speech, and the decoding model was only built from the overt speech data.

i. Decoding overt speech intervals

Results of the decoding of speech and non-speech states for the first patient during overt speech are summarized in **Fig. 78**. Speech intention was decoded using the 41 neural features extracted from only one speech-specific electrode.

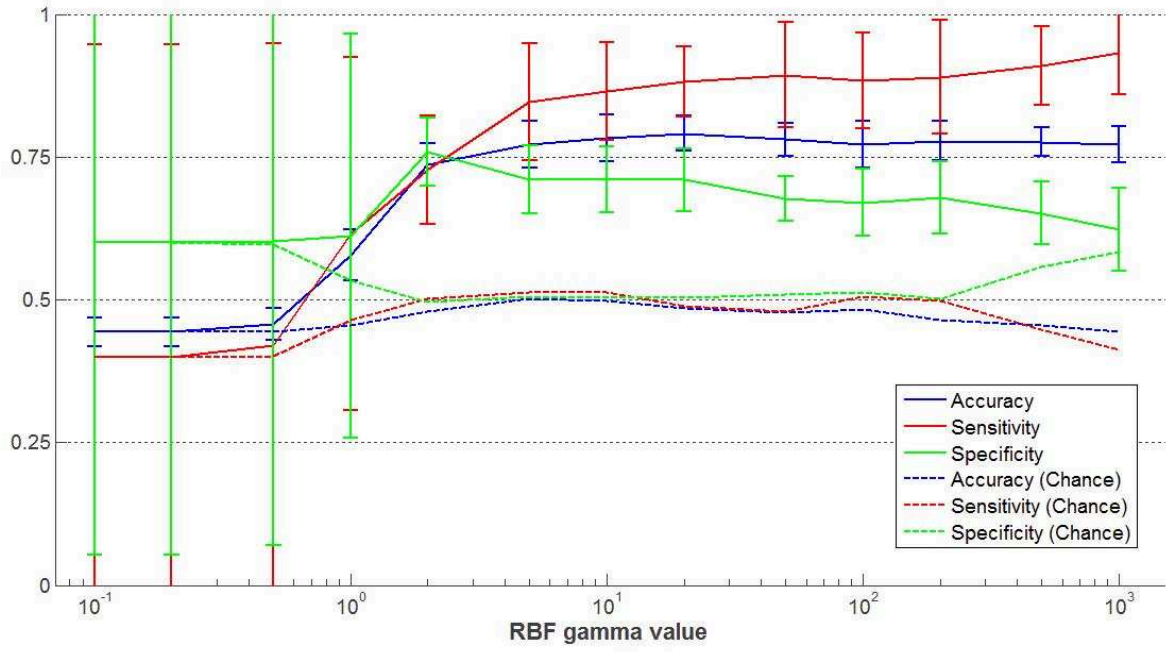


Fig. 78: Decoding of speech and non-speech states for the first patient. The decoding quality was assessed by computing the mean accuracy (blue), the mean sensitivity (red) and the mean specificity (green) over the five cross-validation folds, for different values of the γ parameter of the RBF kernel function. Vertical bars indicate the standard deviation, and dashed lines correspond to chance level.

Average chance levels were 48% for accuracy, 45% for sensitivity and 51% for specificity. Decoding accuracy, sensitivity and specificity are above chance levels for γ values above 1. Best accuracy is achieved for $\gamma = 20$ with a value of 79±3%. For this accuracy, sensitivity is equal to 88±6% and specificity to 71±5%.

ii. Decoding covert speech intervals

The decoding model that was trained on the overt speech data of the first patient was then applied to the covert speech data without further modification. **Fig. 79** shows an example of continuous prediction of the speech intention along with the ground truth, as reported by the patient.

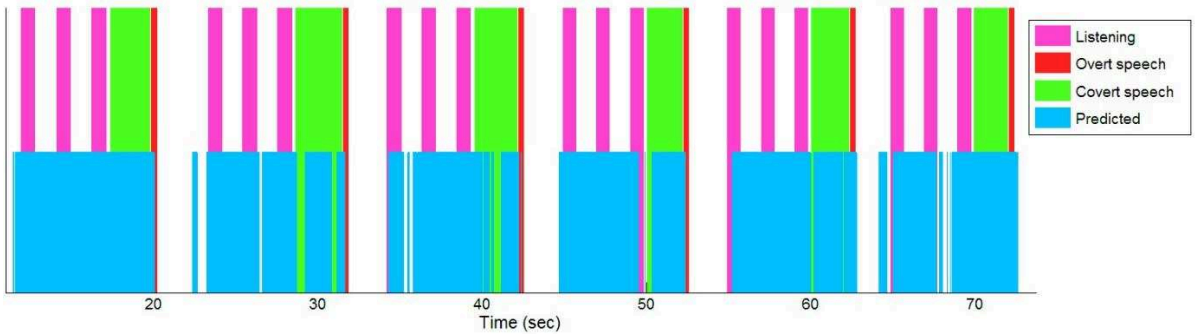


Fig. 79: Speech intention (covert speech) prediction. The patient was first listening to each item presented three times at a fixed pace (pink areas). He was then asked to imagine to repeat them at the same pace (green areas). The end of the imagination *period was notified by the patient by pronouncing "OK" out loud (red areas). The decoding model trained on overt speech* was then applied to this covert speech data without any further modification. The blue areas shows the speech intention predicted by this model.

The decoding model that was only trained on overt speech data was still able to predict speech intention during covert speech periods (see overlapping blue and green areas in **Fig. 79**), with an accuracy of 78% (i.e. 78% of the covert speech periods were actually detected as speech intention). As expected, it also predicted speech intention during actual overt speech (see overlapping blue and red areas in **Fig. 79**). Interestingly, listening periods were also decoded as speech intention in 89% of the listening periods (see overlapping blue and pink areas in **Fig. 79**). In between trials, the model predicted no speech intention (see the gap between each red area and the next pink area in **Fig. 79**), with an accuracy of 84%.

b. Second patient

The second patient only produced overt speech. Results of the decoding of speech and non-speech states from the second patient's data are summarized in **Fig. 80**. Speech intention was decoded using the 120 neural features extracted from 20 speech-specific electrodes (this task being simpler than the decoding of full articulatory trajectories, we did not consider all the different parametrizations of the neural data). There was no delay, nor context added to the neural data, as opposed to when decoding articulatory trajectories.

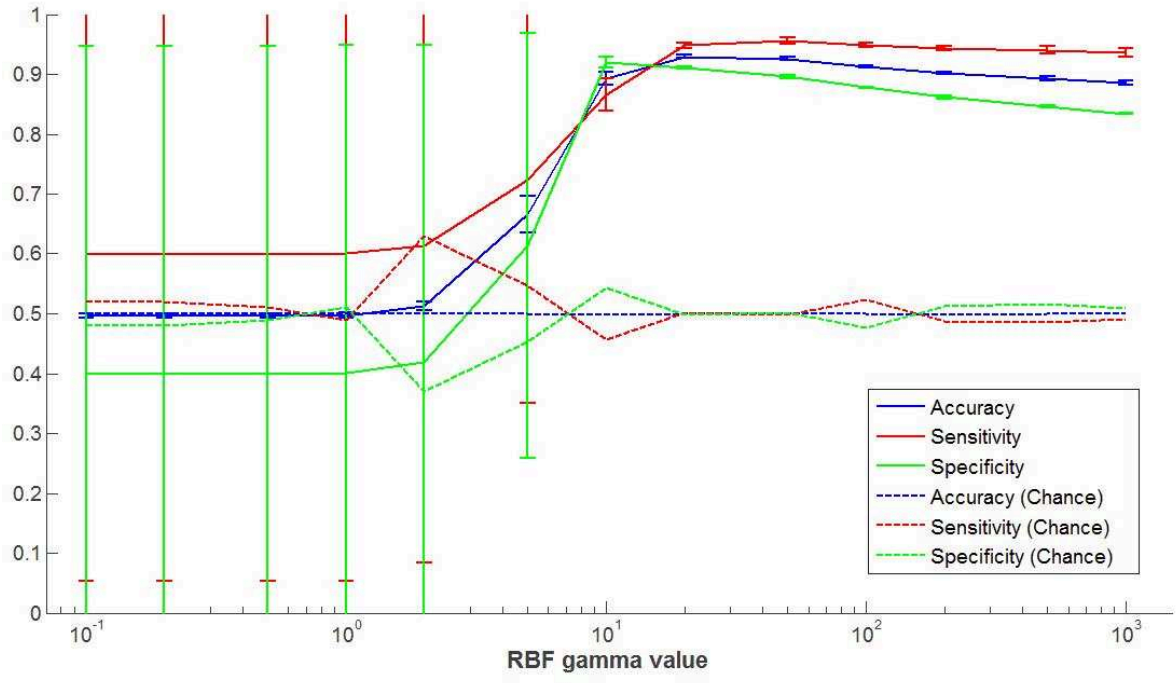


Fig. 80: Decoding of speech and non-speech states for the second patient. The decoding quality was assessed by computing the mean accuracy (blue), the mean sensitivity (red) and the mean specificity (green) over the five cross-validation folds, for *different values of the σ parameter* of the RBF kernel function. Vertical bars indicate the standard deviation, and dashed lines correspond to chance level.

Average chance levels were 50% for accuracy, 53% for sensitivity and 47% for specificity. Decoding accuracy, sensitivity and specificity were above chance levels for γ values above 5. Best accuracy is achieved for $\gamma = 20$ with a value of 93.0±0.2%. For this accuracy, sensitivity is equal to 94.8±0.4% and specificity to 91.1±0.2%.

2. Decoding of the voicing activity

In a second step we considered the decoding of the voicing activity (binary feature for voiced and unvoiced sounds), using the neural data from the second patient. **Fig. 81** summarizes the results of the speech intention decoding.

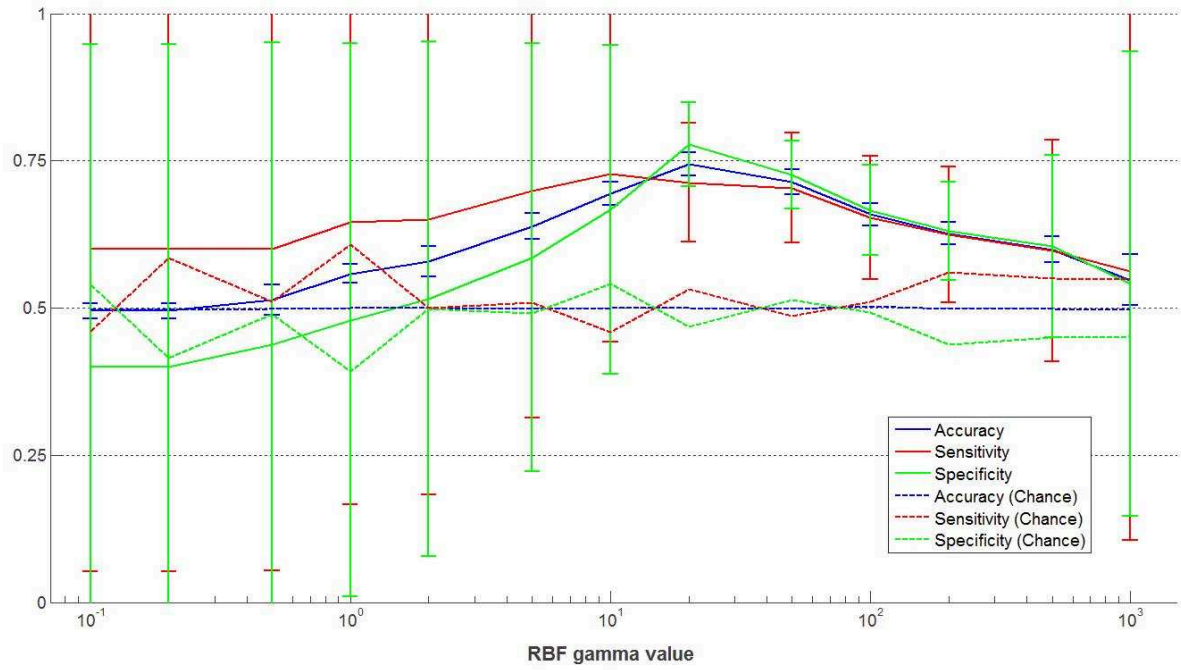


Fig. 81: Decoding of voicing activity. The decoding quality was assessed by computing the mean accuracy (blue), the mean sensitivity (red) and the mean specificity (green) over the five cross-validation folds, for di*fferent values of the σ parameter of* the RBF kernel function. Vertical bars indicate the standard deviation, and dashed lines correspond to chance level.

Average chance levels were 50% for accuracy, 54% for sensitivity and 46% for specificity. Decoding accuracy, sensitivity and specificity were above chance levels for γ values between 5 and 200. Best accuracy was achieved for $\gamma = 20$ with a value of 74±2%. For this accuracy, sensitivity is equal to 71±10% and specificity to 78±7%.

3. Decoding of articulatory features

In a third step we aimed at decoding articulatory trajectories (14 EMA positions) from the neural data of the second patient. The decoding accuracy was assessed by computing the correlation between the predicted and the ground truth articulatory trajectories.

Fig. 82 shows the best correlation for each articulatory feature obtained by spanning all the possible combinations of the neural features parameters (delay, context size and number of PCA components), compared to the best chance level. Thus, the optimal combination was different for each articulatory trajectory.

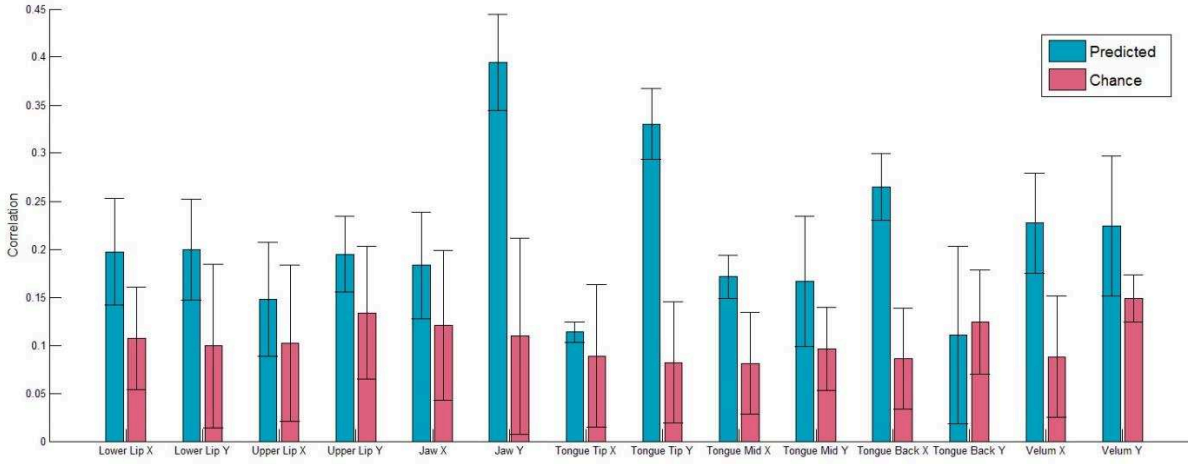
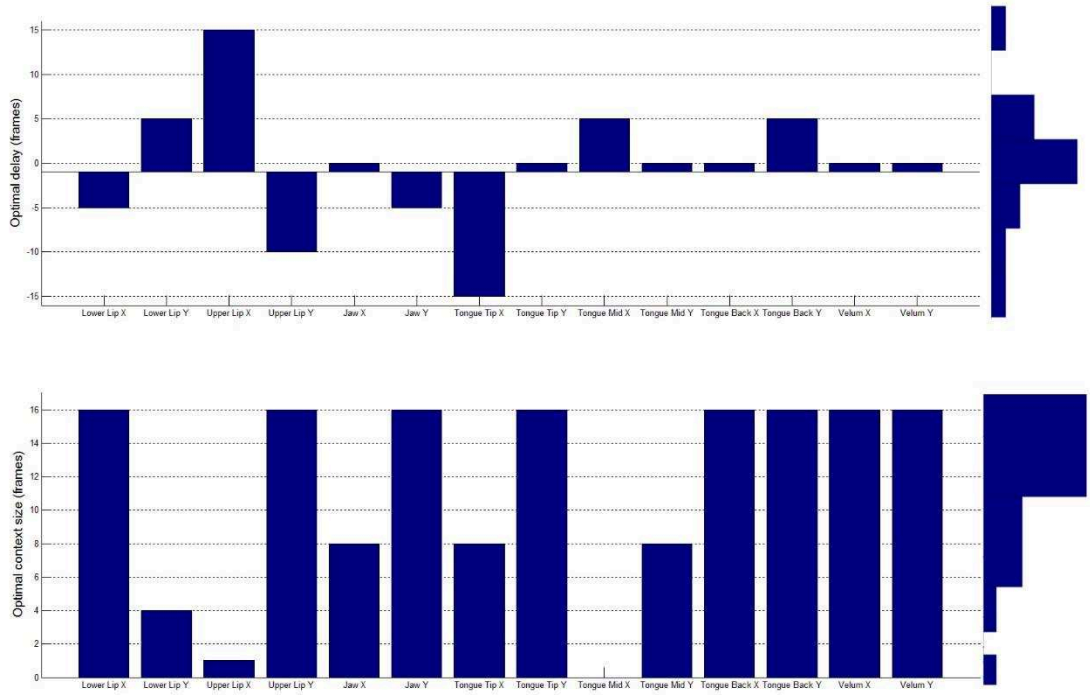


Fig. 82: Best decoding correlation for each articulatory feature. Each bar indicates the best correlation between predicted and reference articulatory features obtained from the neural-to-articulatory mapping (blue) along with the corresponding best chance level (red). Vertical bars indicate the standard deviations.

Decoding performances were systematically above chance level except for the vertical position of the back of the tongue. In particular, significant above-chance decoding could be observed for the vertical position of the jaw (0.39 versus 0.10), the vertical position of the tongue tip (0.33 vs. 0.08), the horizontal position of the back of the tongue (0.26 vs. 0.09) and the horizontal position of the velum (0.23 vs. 0.09).

Fig. 83 shows the optimal delay, context size and number of PCA components for each articulatory feature (i.e. the parameters that led to the best decoding).



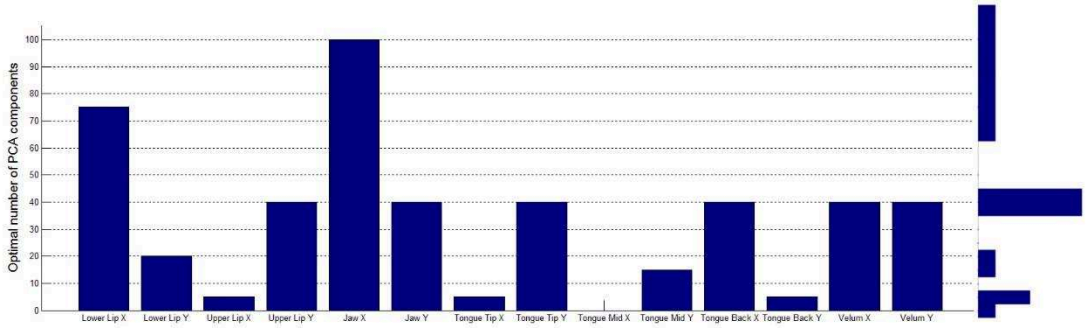


Fig. 83: Optimal delay, context size and number of PCA components for the decoding of each articulatory feature. Top row – Each bar indicates the optimal delay between the neural and the articulatory features (a negative delay meaning that the neural data occurred before the actual speech). Middle row – Each bar indicates the optimal context size for decoding each articulatory feature. Bottom row – Each bar indicates the optimal delay number of PCA components for decoding each articulatory feature. All rows – the bar plot at the right shows the distribution of the parameters values.

The optimal delay varied for each articulatory features but was globally close to 0 for articulatory features decoded above chance level (Jaw Y, Tongue Tip Y, Tongue Back X and Velum X). A similar behaviour could be observed for the optimal context size, which were systematically 16 for more than half the articulatory features including all above-chance features; and similarly the optimal number of PCA components was equal to 40 for all above-chance articulatory features.

Observing the evolution of the decoding accuracy with regards to each parameter (delay, context and number of PCA components) could help identify their individual impact on the decoding quality. **Fig. 84** shows the correlation for the different delays between the neural data and the articulatory features (a negative delay meaning that the neural data occured before the actual speech). **Fig. 85** shows the correlation for the different context sizes (i.e. the number of consecutive neural data frames used). Finally, **Fig. 86** shows the correlation for different number of PCA components kept. In all cases, the displayed correlation was the best cross-validated mean correlation obtained. For instance, the evolution of the correlation with respect to the delay was computed for each delay by keeping the number of PCA components and the context size that led to the best cross-validated mean correlation.

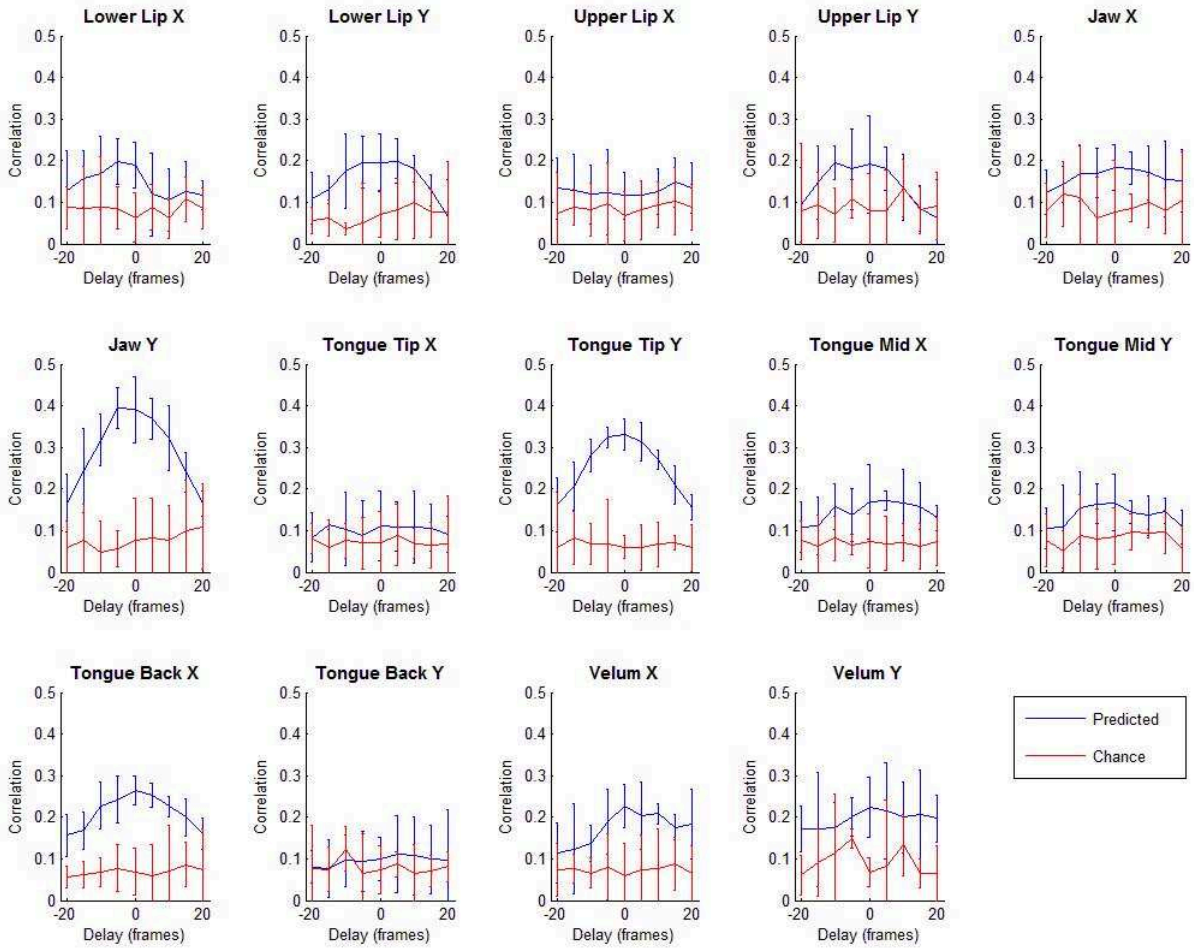


Fig. 84: Decoding accuracy for each articulatory parameter with respect to the delay between neural and articulatory data. Each plot shows the mean correlation between the predicted and ground truth values (blue line), as well as chance level (red line), for each articulator and each delay. The delays are in data frames (1 frame = 10ms). A negative delay means that the neural data was considered before the actual speech. Vertical bars correspond to standard deviations.

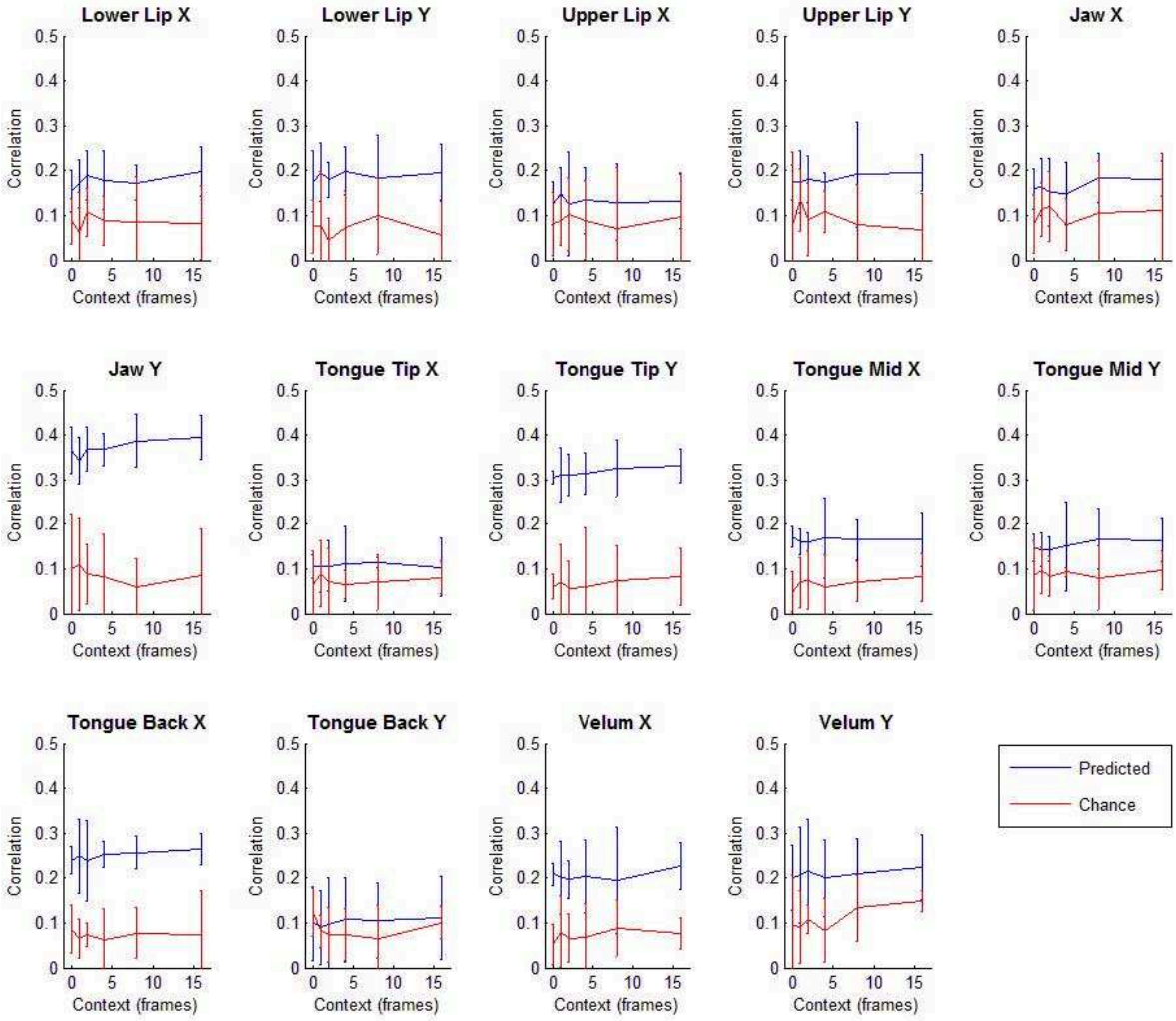


Fig. 85: Decoding accuracy for each articulatory parameter with respect to the context size. Each plot shows the mean correlation between the predicted and ground truth values (blue line), as well as chance level (red line), for each articulator and each context size. The context sizes are in data frames (1 frame = 10ms). Vertical bars correspond to standard deviations.

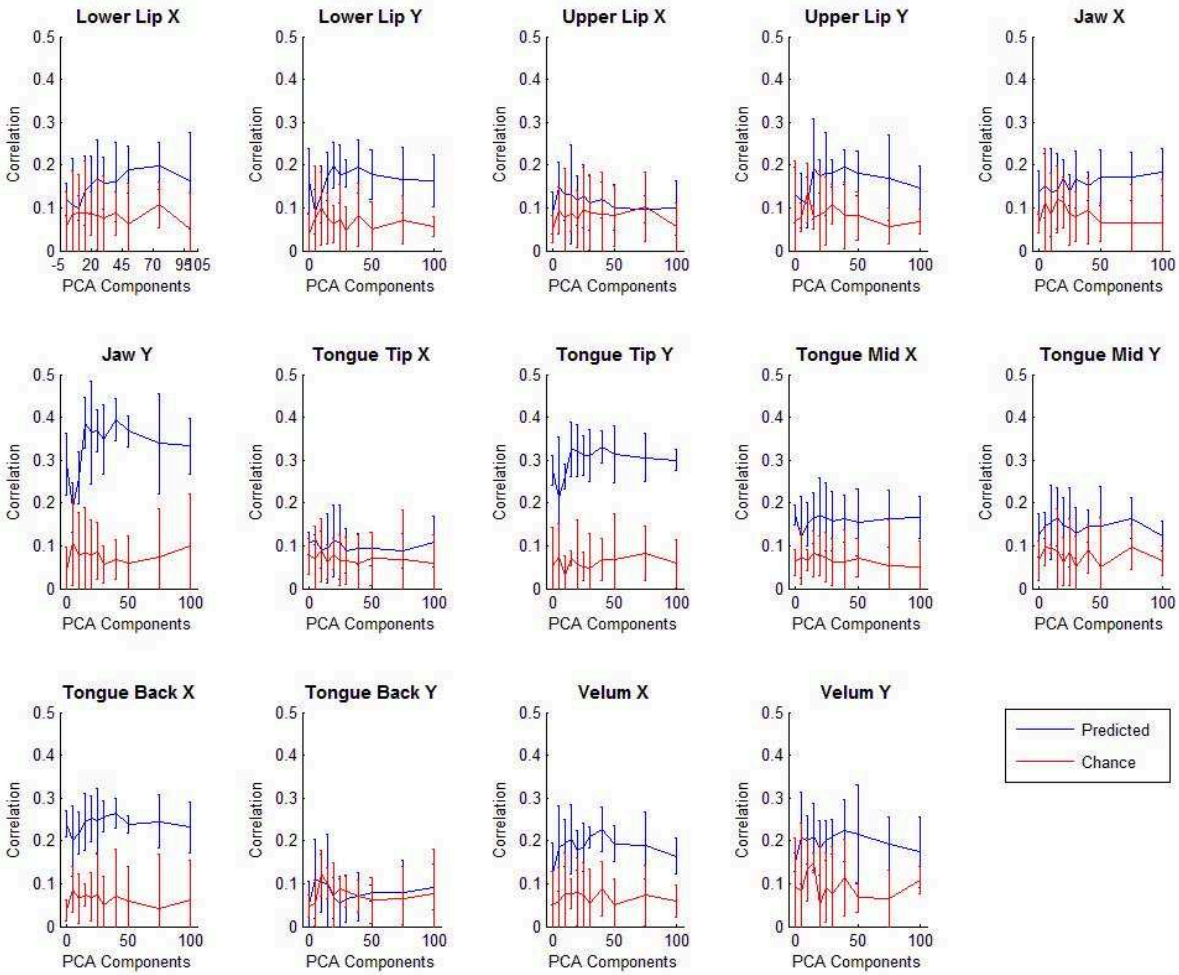


Fig. 86: Decoding accuracy for each articulatory parameter with respect to the number of PCA components. Each plot shows the mean correlation between the predicted and ground truth values (blue line), as well as chance level (red line), for each articulator and each PCA components number. A number of 0 corresponds to not using PCA. Vertical bars correspond to standard deviations.

Overall, mean predicted articulatory features were above chance level. The delay between the neural and the articulatory features clearly had an impact on several articulators, especially for the vertical position of the jaw, tongue tip and lower lip, as well as the horizontal position of the tongue back, with best correlation achieved when there was no delay (**Fig. 84**). Increasing the context size generally resulted in a slight increase of the decoding accuracy (**Fig. 85**). Such direct relationship between the number of PCA components and the decoding quality was not observed. However, using PCA globally lead to best correlation than when not using PCA (here indicted by a null number of components), and the optimal number of principal components was generally between 20 and 50 (**Fig. 86**).

4. Decoding of acoustic features

In order to evaluate the benefits of using articulatory trajectories as an intermediate representation when decoding speech from neural activity recorded in the motor cortex, we also performed the decoding of acoustic features (25 mel coefficients) for the same patient by using the original audio recorded during the experiment. The decoding accuracy was assessed by

computing the the correlation between the predicted and the ground truth acoustic trajectories 25 mel coefficients).

Fig. 87 shows the best correlation for each acoustic feature obtained by spanning all the possible combinations of the neural features parameters (delay, context size and number of PCA components), compared to the best chance level. Thus, the optimal decoded was different for each mel coefficient.

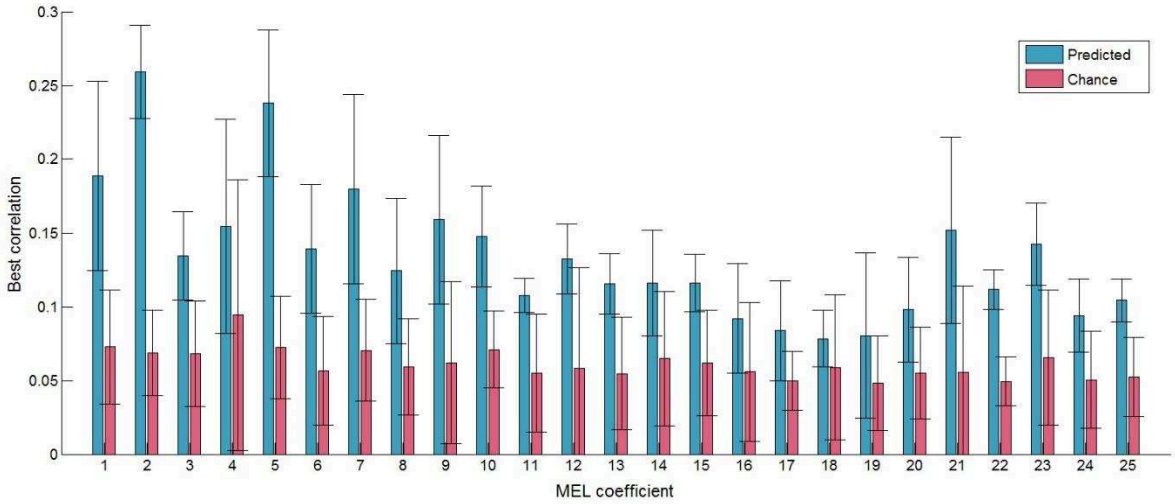


Fig. 87: Best decoding correlation for each acoustic feature. Each bar indicates the best correlation between predicted and reference mel coefficients obtained from the neural-to-articulatory mapping (blue) along with the corresponding best chance level (red). Vertical bars indicate the standard deviations.

Results were systematically above chance level and significant above-chance decoding can be observed for the first (0.19 versus 0.07), second (0.26 vs. 0.07), fifth (0.24 vs. 0.07) and seventh (0.18 vs. 0.07) MEL coefficients.

Fig. 88 shows the optimal delay, context size and number of PCA components for each acoustic feature (i.e. the parameters that led to the best decoding).

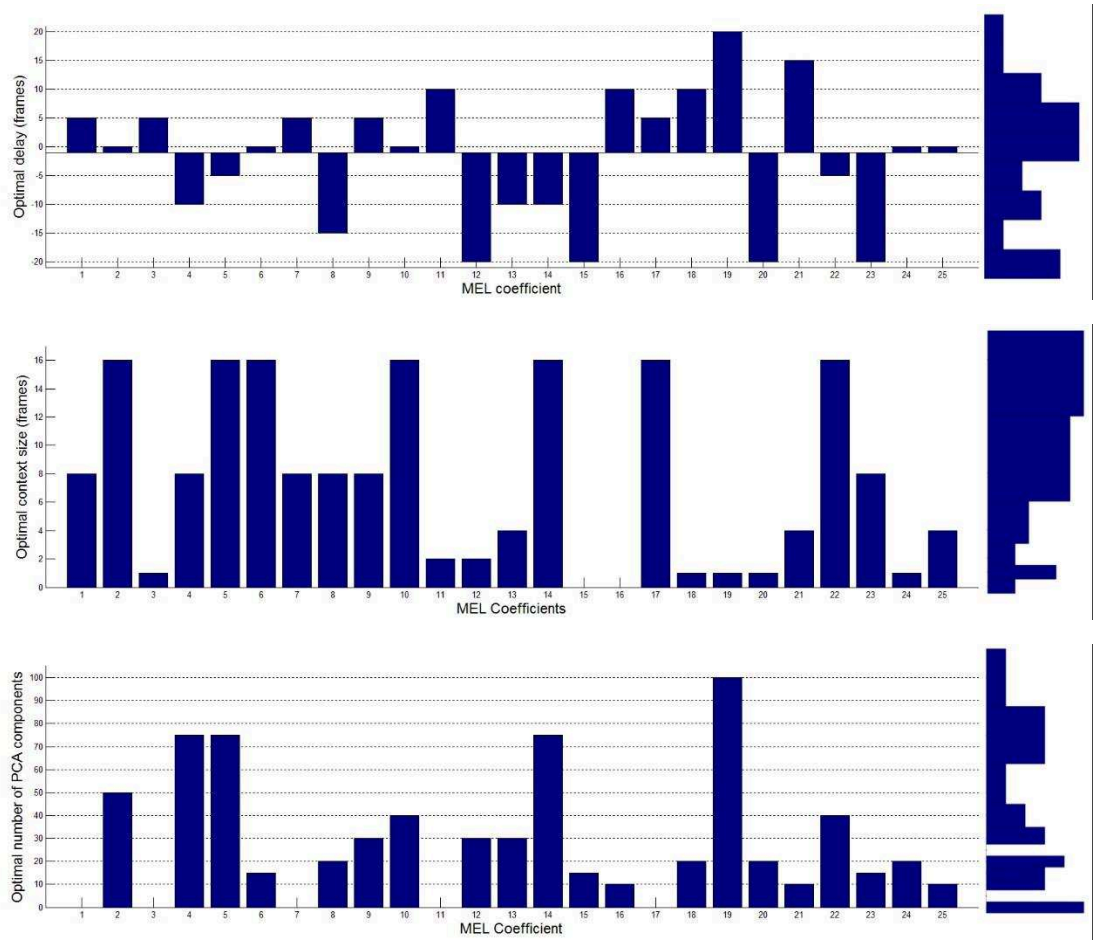


Fig. 88: Optimal delay, context size and number of PCA components for the decoding of each acoustic feature. Top row – Each bar indicates the optimal delay between the neural and the acoustic features (a negative delay meaning that the neural data occurred before the actual speech). Middle row – Each bar indicates the optimal context size for decoding each acoustic feature. Bottom row – Each bar indicates the optimal delay number of PCA components for decoding each acoustic feature. All rows – the bar plot at the right shows the distribution of the parameters values.

As for articulatory features, the optimal delay varied but was globally close to 0 for articulatory features decoded above chance level (first, second, fifth and seventh MEL coefficients). Most optimal context sizes were above 8 frames (i.e. more than 80ms of neural data). No consistent behaviour could be observed for the optimal number of PCA components, which varied from 0 (no PCA) to 100.

The evolution of the decoding accuracy with regards to each parameter (delay, context and number of PCA components) was then observed to try to identify their individual impact on the decoding quality. **Fig. 89** shows the correlation for the different delays between the neural data and the acoustic features (a negative delay meaning that the neural data occured before the actual speech). **Fig. 90** shows the correlation for the different context sizes (i.e. the number of consecutive neural data frames used). Finally, **Fig. 91** shows the correlation for different number of PCA components kept. In all cases, the displayed correlation was the best cross-validated mean correlation obtained. For instance, the evolution of the correlation with respect to the delay was computed for each delay by keeping the number of PCA components and the context size that led to the best cross-validated mean correlation.

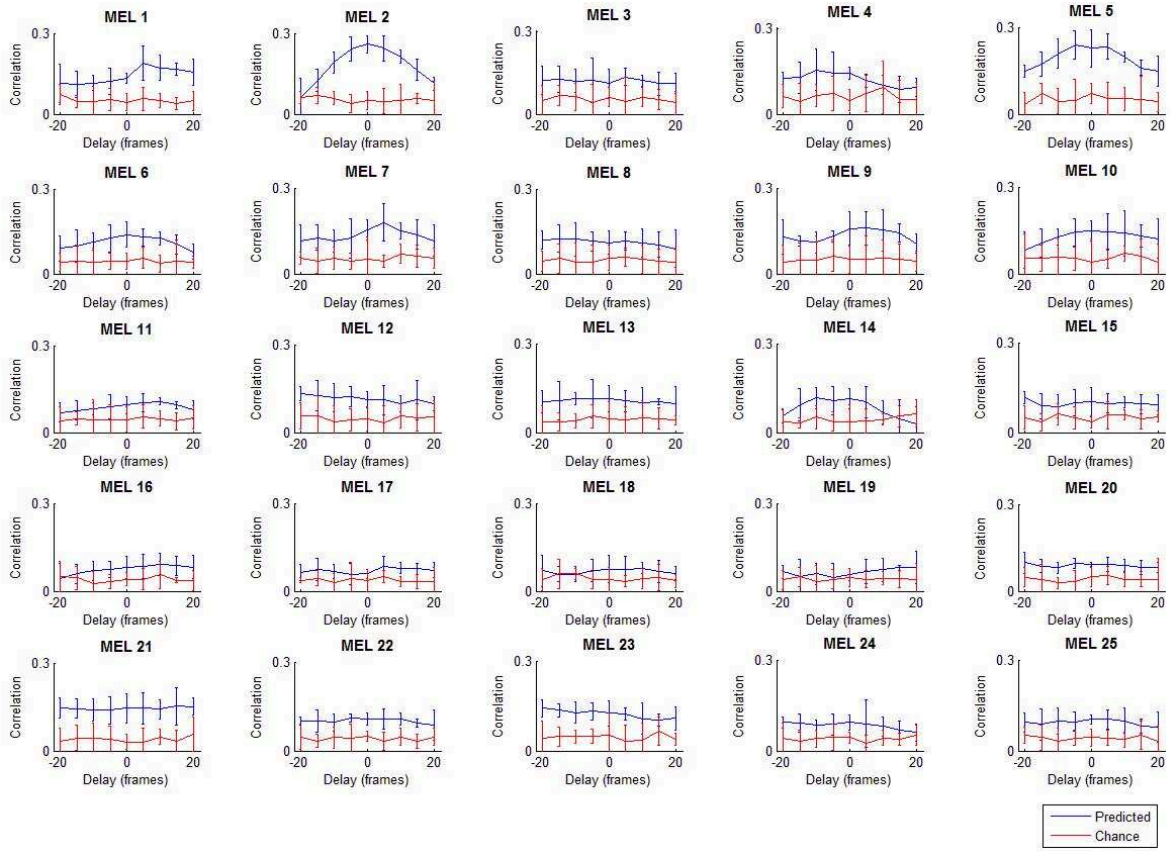


Fig. 89: Decoding accuracy for each acoustic parameter with respect to the delay between neural and acoustic data. Each plot shows the mean correlation between the predicted and ground truth values (blue line), as well as chance level (red line), for each mel coefficient and each delay. The delays are in data frames (1 frame = 10ms). A negative delay means that the neural data was considered before the actual speech. Vertical bars correspond to standard deviations.

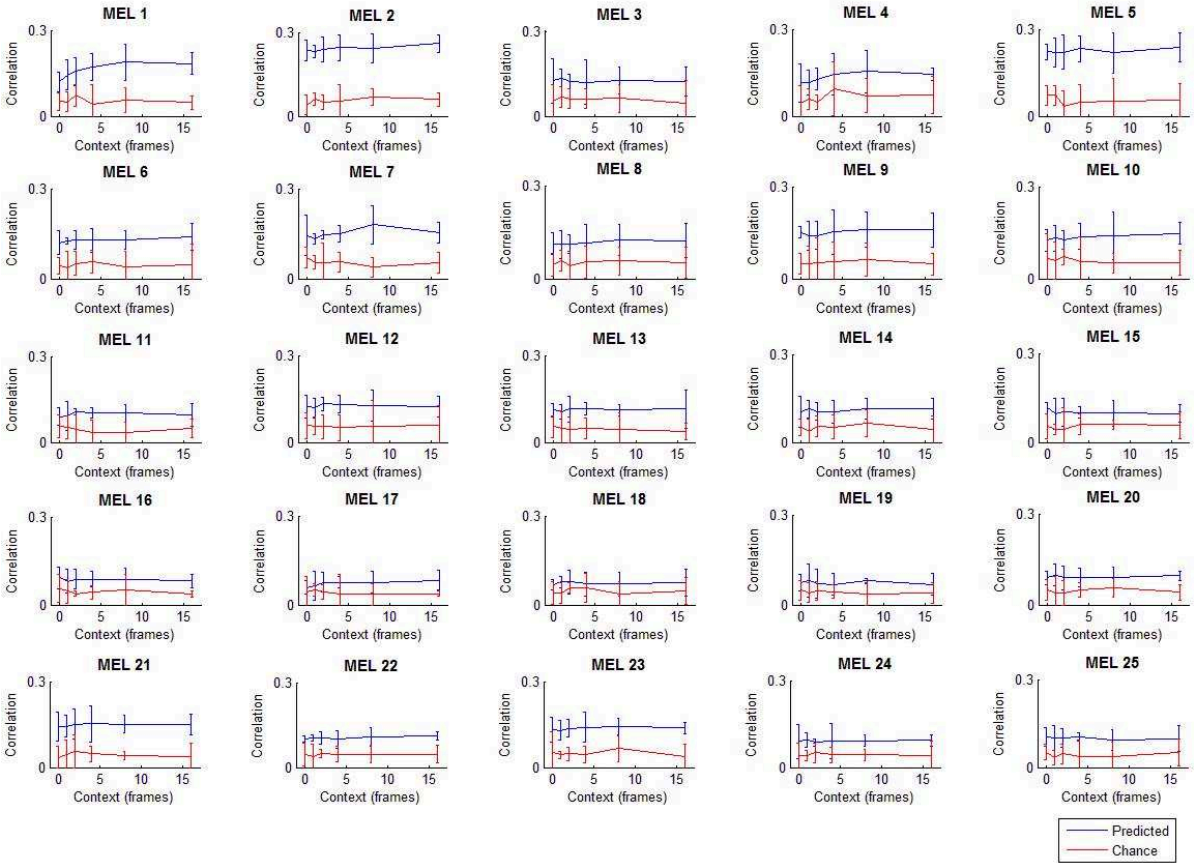


Fig. 90: Decoding accuracy for each acoustic parameter with respect to the context size. Each plot shows the mean correlation between the predicted and ground truth values (blue line), as well as chance level (red line), for each mel coefficient and each context size. The context sizes are in data frames (1 frame = 10ms). Vertical bars correspond to standard deviations.

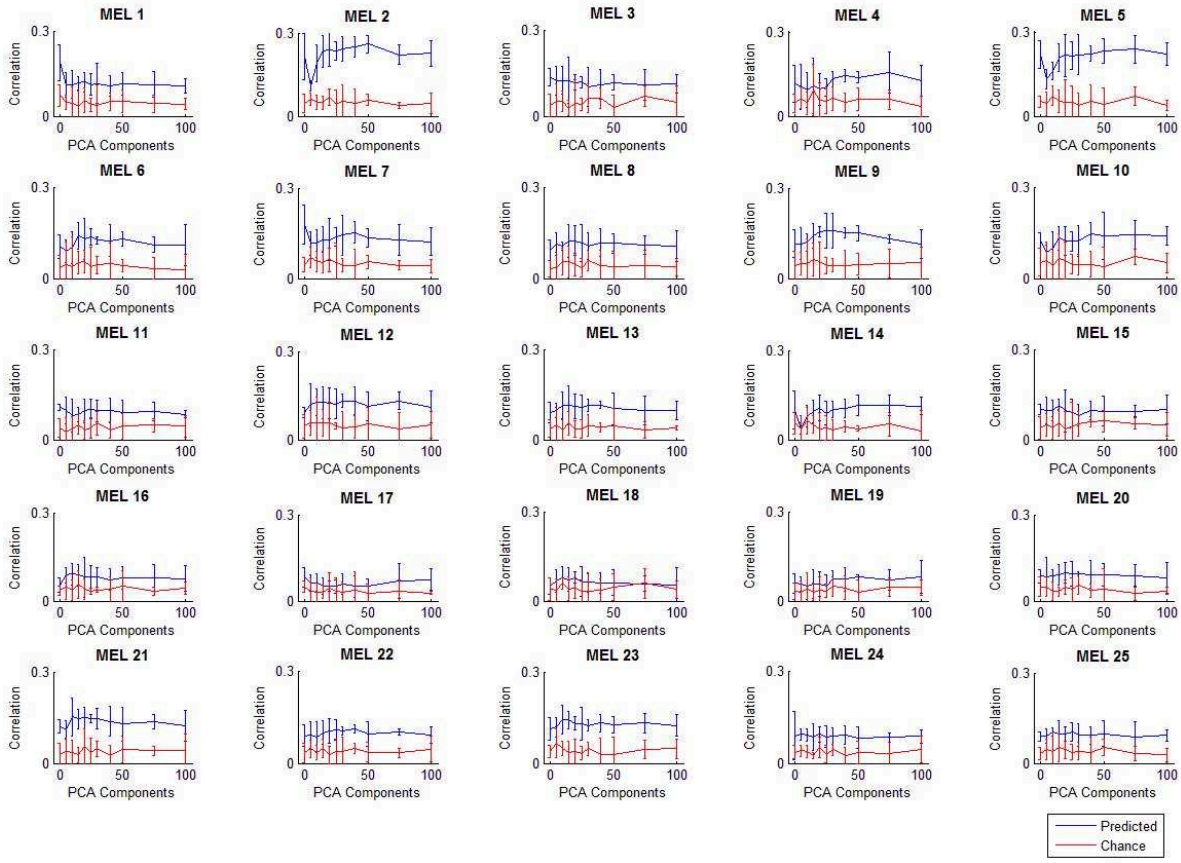


Fig. 91: Decoding accuracy for each acoustic parameter with respect to the number of PCA components. Each plot shows the mean correlation between the predicted and ground truth values (blue line), as well as chance level (red line), for each mel coefficient and each PCA components number. A number of 0 corresponds to not using PCA. Vertical bars correspond to standard deviations.

Overall, mean predicted acoustic features were systematically above chance level. The delay between the neural and the acoustic features principally impacted the second and fifth mel coefficients, with best decoding achieved when there was no delay (**Fig. 89**). Increasing the context size generally resulted in a slight increase of the decoding accuracy, especially for the first MEL coefficient, for which the decoding correlation continuously increased from 0.12 without context to 0.19 with an 8-frames context (**Fig. 90**). Such direct relationship between the number of PCA components and the decoding quality was not observed (**Fig. 91**).

5. Comparison of the neural-to-articulatory and neural-to-acoustic mappings

In order to compare the neural-to-articulatory and the neural-to-acoustic decoding, the predicted articulatory features were mapped to acoustic features using the articulatory-to-acoustic mapping of the articulatory-based synthesis (see **Methods**). The acoustic features obtained through the neural-to-articulatory decoding were compared to the aligned reference BY2014 acoustic features, while the acoustic features obtained from the neural-to-acoustic decoding were compared to the reference patient's acoustic features. **Fig. 92** shows an example of predicted and reference patient's acoustic features (bottom row), as well as the corresponding predicted and reference BY2014 acoustic features (top row), for the second mel coefficient (which is the best decoded acoustic feature).

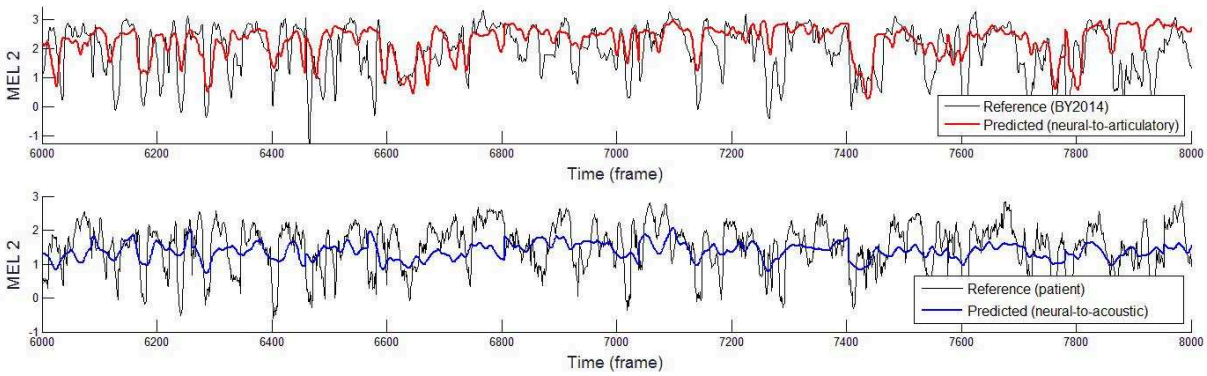


Fig. 92: Example of predicted and reference acoustic features. Top row – BY2014 acoustic features (black) and the predicted acoustic features using the neural-to-articulatory and the articulatory-to-acoustic mappings (red). Bottom row – Reference *patient's acoustic features (black) and the predicted acoustic features using the neural*-to-acoustic mapping (blue).

In this example, the reconstruction of the second mel coefficient through the neural-to-articulatory decoding was qualitatively better than when using the neural-to-acoustic decoding. We further quantified this systematically. To compare both mapping approaches, the correlation between the predicted and appropriate reference acoustic features for each mapping was computed (**Fig. 93**). For this comparison, the delay, context size and number of PCA components were fixed to those that led to the best decoding (see **Methods** and previous sections). For the neural-to-articulatory mapping, this resulted in having a negative delay of 5 frames (i.e. taking neural activity 50ms in the past), a context size of 16 frames (i.e. considering 160ms continuous data chunks) and 40 PCA components. For the neural-to-acoustic mapping, there was no delay, a context size of 16 frames as well (i.e. considering 160ms data chunks in the past), and 50 PCA components.

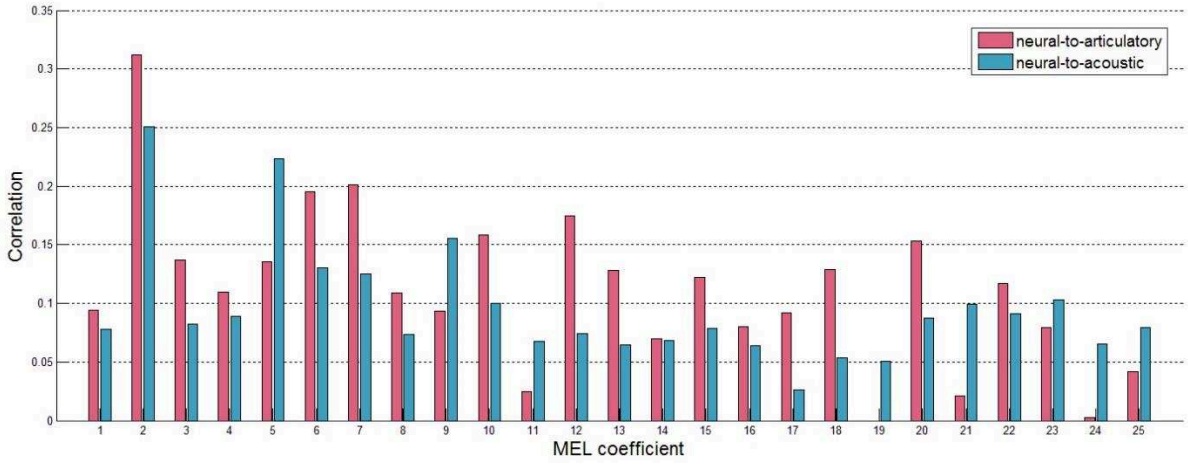


Fig. 93: Comparison of the neural-to-acoustic and neural-to-articulatory mapping. Each bar shows the correlation between predicted and reference mel coefficients fot the neural-to-acoustic (blue) and neural-to-articulatory (red) mappings.

For 17 out the 25 MEL coefficients, the neural-to-articulatory mapping resulted in a higher correlation with the reference acoustic features than for the neural-to-acoustic mapping. In both cases, the decoding accuracy was not high enough to synthesize intelligible speech.

IV. Conclusion on the speech decoding from neural activity

In this chapter we presented three different decoding steps required to build a BCI for speech rehabilitation based on articulatory data.

First, we considered speech intention detection, i.e. the prediction of the intention of speaking from neural data. The decoding was performed using support vector machines (SVM) on neural data recorded during brain surgery in two patients. For both patient, the decoding accuracy of speech segments during actual speech production (overt speech) was far above chance level (50%) with 79% accuracy for the first patient (**Fig. 78**), and 93% accuracy for the second one (**Fig. 80**). The higher accuracy for the second patient might essentially come from the fact that more and smaller electrodes were used than for the first patient, resulting in more electrodes with speech-specific neural activity (20 electrodes for the second patient versus 1 for the first one). This suggests that the decoding accuracy could still increase by having more electrodes covering the speech areas. This was consistent with results observed in most BCI studies, showing that the decoding accuracy generally increases when increasing the number of electrodes (Carmena et al., 2003; Pasley et al., 2012). Then, we also considered speech intention detection for imagined speech (covert speech). Only the first patient had data recorded during covert speech. For this patient, we used the SVM previously trained on the overt speech data to decode covert speech segments. Results showed that speech intention during covert speech could also be predicted with an accuracy of 78% (**Fig. 79**). This accuracy is very similar to that of decoding overt speech (79%). Moreover, speech intention was also predicted during listening periods (89%), which is not due to a poor specificity of the decoding model since it was able to correctly predict rest period with an accuracy of 84%. This could explain the limited accuracy of the speech intention decoding during overt speech. Indeed, listening periods in the overt speech data were labeled as non-speech states but could have been decoded as speech intention, resulting in a drop of accuracy. Moreover, these results suggest that the neural activities during speech production, speech imagination and speech listening share at least a partially common representation in the speech motor cortex. This is consistent with the mapping of speech-related activity that exhibited common features between overt and covert speech (see **Chapter 6**). These three different tasks could eventually be distinguished by using denser recordings, as it was the case for the second patient.

In a second step, we considered voicing activity decoding, i.e. predicting if the vocal folds are vibrating or not during speech. Only the data from the second patient, and during speech segments, was considered. Results showed that the voicing activity could be decoded with an accuracy of 74%, which is above chance level (**Fig. 81**). However, the decoding accuracy remains relatively low compared to the speech intention detection (93%). This might be caused by the fact that electrodes did not cover enough the speech motor cortex, and could thus not capture enough information to decode the voicing activity with higher accuracy. Future studies should consider a better coverage of the speech motor cortex.

Finally, we considered the decoding of articulatory trajectories from the neural data of the second patient. The decoding of the articulatory trajectories was achieved using linear models, since the number of samples was too limited to benefit from more advanced techniques. Results showed that several articulatory features could be decoded above chance level, in particular the vertical position of the jaw (correlation of 0.39), the vertical position of the tongue tip (0.33),

the horizontal position of the back of the tongue (0.26) and the horizontal position of the velum (**Fig. 82**). Several parameterization of the neural data were tested, by varying the delay between the neural and articulatory data, the duration of each neural data frame, and the number of dimensions of the neural data. Results were variable among the different articulatory features but suggested that best accuracy could be achieved without any delay between the neural and articulatory data (**Fig. 84**). This is consistent with the results obtained from the functional mapping of speech-related brain activity that exhibited an increase in low and high-gamma bands during and not before the production of speech (see **Chapter 6**). Best decoding accuracy was also reached when considering neural data chunks of at least 100ms, and results suggest that increasing the context size even more could results in improving the decoding accuracy (**Fig. 85**). Future works should thus consider longer chunks of neural data for the decoding of articulatory trajectories. Moreover, results showed that reducing the dimensionality of the neural feature space could improve the decoding accuracy (**Fig. 86**). While the decoding accuracy remained too limited to generate intelligible speech, it showed that the speech motor cortex encodes at least a part of the articulatory properties of speech during speech production, which is consistent with other studies (Bouchard et al., 2013; Cheung et al., 2016).

To test the hypothesis that decoding articulatory trajectories as input to an articulatory-based speech synthesizer could be more efficient that directly predicting acoustic parameters to synthesize speech, we as well considered the decoding of acoustic features from neural activity. The decoding of acoustic trajectories was performed in the same way as for articulatory trajectories, except that MEL coefficients were used instead of EMA data. Results showed that several MEL coefficients could be decoded above chance level, in particular the first (correlation of 0.19), second (0.26), fifth (0.24) and seventh (0.18) coefficients (**Fig. 87**). As for the decoding of articulatory trajectories, results suggested that best accuracy could be achieved without any delay between the neural and acoustic data (**Fig. 84**),when considering neural data chunks of at least 100ms (**Fig. 85**), and when using dimensionality reduction techniques (**Fig. 86**). This suggests that the speech motor cortex could encode at least a part of the acoustic properties of speech during speech production. However, it is known that acoustic and articulatory parameters are correlated (Bouchard et al., 2016), so that further analysis should be performed to determine whether the speech motor cortex encode for articulatory trajectories, acoustic features, both or another representation.

To investigate this aspect, the decoded acoustic features were compared depending on whether they were decoded directly from neural data or from articulatory trajectories estimated from neural data. Results showed that most of the acoustic features were best decoded when passing by the intermediate articulatory representation (**Fig. 93**). This is particularly interesting given that the articulatory trajectories by a model trained from articulatory data of another subject (from the BY2014 dataset), aligned using the recorded patient's audio. Indeed, the audio of the BY2014 dataset was aligned on the patient's audio using dynamic time warping, which resulted in a good but not perfect alignment, and this time warping was then applied to articulatory trajectories from the BY2014 dataset that were not recorded from the patient himself. This could have resulted in a mismatch between the patient's neural activity and the target articulatory trajectories used to train the decoding model. However, these preliminary results must be confirmed by future studies.

Altogether, these results are first steps toward a BCI for speech restoration using articulatory data as an intermediate representation of speech. However, further experiments are

needed to confirm these preliminary results. In particular, future works should focus on the decoding of covert speech from the speech motor cortex.

Part 5: Thesis result 3 – Ethical aspects

During my thesis, we had the chance to work with Nicolas Aumonier on some ethical aspects of brain-computer interfaces. While ethics were not part of this thesis subject, I personally think that any researcher should consider ethical aspects of his work. In this part, I will briefly present some of the major ethical issues that arise when developping brain-computer interfaces (BCIs) in general, and not specifically for speech BCIs. This reflexion led us to the writing of a book chapter (Bocquelet et al., 2016c), which I will try to resume in the following.

The global purpose of brain-computer interfaces (BCIs) is to interface our central nervous system (CNS) with external devices in order to restore lost functions following a disease or an accident. Since the CNS is generally considered as the most intimate seat of thought, consciousness and personality, several ethical issues arise. Who can benefit from these technologies and for what purpose? Where should be the border between rehabilitation and enhancement? How to manage the hope aroused in patients participating to clinical developments? In the long term, can we identify safety, security and legal issues raised by the usages of BCIs? More fundamentally, could BCIs modify the identity of a person? Could they lead to a redefinition of humanity?

While most of these questions will remain unanswered at the end to this chapter, the objective here is to raise them and to bear them in mind while conducting research.

I. Introduction

BCIs are an emerging technology that aims to establish a direct communication between the brain and a computer or a machine in order to directly control effectors, objects or software from thoughts. These approaches thus offer a new interaction mode between living individuals and machines, which contrast with those commonly used today such as touch or vocal modes.

With neuroprosthetics, the objective of BCIs is to interface the CNS with arrays of electrodes or sensors allowing recording of neuronal activity and/or delivering controlled electrical stimulations to trigger activities to recover of a function. Some of these system can be non-invasived – for instance electroencephalography, magnetoencephalography or even functional magnetic resonance imaging – and other are invasive and require the electrodes in the CNS.

These technologies are developed for multiple purposes: clinical applications that propose therapeutic routes in the case of sensory or motor disabilities or neurodegenerative diseases such as in (Benabid et al., 1991; Margalit et al., 2002; Hochberg et al., 2006; Guenther et al., 2009), fundamental research employing these technologies as a means to explore in a novel manner the functioning of the CNS (Fetz, 1969, 2007; Carmena et al., 2003; Jackson et al., 2006; Lachaux et al., 2007; Moritz et al., 2008; Ganguly and Carmena, 2009; Engelhard et al., 2013; Mercier-ganady et al., 2014), and all other application such as military applications (Tennison and Moreno, 2012) or consumer application such as entertainment (Congedo et al., 2011; Bonnet et al., 2013).

It seems now quite likely that, within several years, these technologies will have more and more applications, that they will become available on the market and thus impact our society. Although the consequences of these technologies on our society remain difficult to predict, several questions can be formulated on the ethical, legal, political, economic, philosophical, moral and religious levels. Several ethical argumentation methods are available and a large number of authors compete against each others. Nowadays, whereas in applied ethics the discussion may focus on what is good or not, in moral philosophy these rather focus on the arguments that make it possible to justify an ethical choice than on this choice itself (Canto-Sperber and Ogien, 2004).

An agent may indeed want to justify its action by referring to the good, the duty or the usefulness. Aristotle bounded the good to happiness so that an agent have good chances to be happy if his actions are directed toward the good, which is supposed to exist objectively (Aristote, n.d.). However, if the good is not considered as objective, and so happiness, it might be necessary to rely on a more impartial judge that we can call duty (Kant, 1785). Good will then be the will that acts under the influence of no particular interest, but by pure duty, which everyone perceives identically through their inner consciousness. However, if such objective duty makes it difficult to find a concrete direction, then we might want to turn toward what experienced people deem most useful, to what we like most and affects us the least (Mill, 1863). In that case, the agent will seek to maximize for usefulness for the ethical agents and to minimize everything that may cause uselessness or pain. He will thus aim to minimize a risk/benefits ratio. The considered ethical agents can be himself, or those that surround him (family, friends, etc.), all human beings, all sentients beings (human beings and animals), all

living beings, all ecosystems or our whole planet. For BCIs, it is possible to consider three subjects of ethical questionings, dealing respectively with the animal, human being, and human species. We will try to address at these three levels the various families of BCI technologies: rehabilitation and care solutions for disabled patients, advancement of fundamental research in the understanding of brain mechanisms and the development of new markets including consumer products based on more or less invasive BCIs.

II. The animal

As BCIs typically require preclinical animal experimentations, research on animals are for many researchers the occasion for a real ethical questioning.

1. The fight against pain, suffering and anxiety in animals

The European legislation considers as a value for all European Union (EU) member countries to minimize as much as possible animal pain, suffering and anxiety. In nowadays state-of-the-art laboratories animal facility officials and researchers experimenting with animals put extreme care on the welfare of animals, in a striking ressemblance with the respect shown toward a patient or a human subject: all practical actions on animals are traced, the operating rooms for animals are very similar to an operating room for human beings, and all trials on animals must be submitted to an ethics committee of animal experimentation through application files sometimes thicker than that of applications submitted to committee for the protection of people, which oversees any experimentation on human beings in France. This parallelism is in line with the idea that humans and animals are all beings capable of feeling suffering, which should be minimized as much as possible, as stated in the new directive of the European Commission[2].

2. Animals are not things

Since the suffering of animals has been taken into consideration at the same level as that of human beings, the difference between the human being and the animal has no longer the evidence that it had formerly. Thus, the researcher prepares himself everyday a little more in order to be able to justify, to himself, to the public, and to eventual militants of the animal cause, more or less violent, that it is not possible to expect care or knowledge – knowledge to care, or knowledge to understand – without carrying out trials on animals.

Nevertheless, this does not prevent that a few borderline cases be distinguished. For example, cockroaches have been recently used as a simple game in which they are controlled to move in any direction using a wireless device that is directly connected to their antennae[3]. In

[2] Directive 2010/63 of the European Commission enjoins all member countries to restrain the most possible pain, suffering and anguish in animals upon which scientific experiments are carried out, recommendation translated in France by the decree No. 2013-118 of February 1, 2013 relatively to the protection of animals used for scientific purposes, amending the rural code.

[3] Backyard Brains: The RoboRoach.

this case, human domination is exercised for the simple leisure industry which, in an ethically questionable manner, transforms animals into toys. What usefulness then justifies this application of BCIs? Is the usefulness relying on economic and commercial factors relevant in this case? What limits should be enforced to these approaches that nowadays concern cockroaches or rats (Talwar et al., 2002), but that nothing prevents thinking that they may one day concern human beings? Since many evidences point out that animals are intelligent and sentient beings (Bekoff, 2000; Paul et al., 2005; Reznikova, 2007), which was recently officially recognized in France[4], they should not feel pain or suffering or anxiety when manipulated, and their condition must be respected.

III. The human being

Just like animals, human beings are directly concerned by BCIs, which can have a major impact on their well-being. Numerous ethical questionings emerge when considering the use of BCIs, be they invasive or not, and this is particularly so because their applications are still at the research stage, still immature. These questionings concern in the first place patients who agree to participate to BCI research protocols.

1. Addressing the aroused hope

Indeed, BCI studies often suggest real prospects for improvement among heavily handicapped patients, for whom there is no other alternative at the moment. This sole word of improvement may give rise to many expectancies, some of them being totally unrealistic to immediately benefit the patient, but rather future patients. Thus, a wide gap may exist between the expectation of the patients and what can be offered to them in return of being included in a study designed to test a paradigm, a material, a hypothesis, not aiming at improving the status of a real person (Lidz et al., n.d.; Clausen, 2009). For health professionals, the challenge posed by this gap consists of knowing how to address the hope risen by these new technologies among "expecting" patients, that is in individuals whose state is so serious that they have reached the point to expect everything from medicine, the progress of which is now considered so important that patients tend to put on practitioners the hopes that they once placed in natural or in God's healing: Today doctors are supposed to know, they are supposed to have the power to give patients their health back. Addressing this hope becomes even more significant when the effectiveness of non-invasive BCI systems remains limited, and that a choice has to be made to resort or not to a surgery for testing a more invasive system. Although the present results obtained with invasive interfaces offer extremely promising perspectives with a growing number of degrees of freedom that can be controlled simultaneously (Hochberg et al., 2012; Ifft et al., 2013; Wodlinger et al., 2014), they nevertheless still remain modest (Dietrich et al., 2010) and great importance is still given to more conventional techniques for the compensation of handicaps using interfaces based on residual movements (Pino et al., 2003; Brunner et al., 2010; Treder and Blankertz, 2010; Takahashi et al., 2011). However, these approaches are themselves

[4] The French National Assembly has recognized on October 30, 2014, that animals were "sentient living beings", alining thus the civil code with the penal code and the rural code; the amendment was rejected by the Senate on January 22, 2015, and then confirmed by the National Assembly on January 28, 2015.

limited, and the hypothesis currently favored through the development of new generations of BCI systems is that direct interfacing of the SNC in the long run should offer better rehabilitation prospects.

2. Risk/benefits ratio

Any BCI application must be conducted in accordance with the principles of the patients' autonomy, of their informed consent, as well as the practitioners' commitment about the beneficence of their acts, sworn during the oath to respect the code of medical ethics[5,6]. In the case of invasive BCIs, for which it is necessary to proceed with a surgical implantation of electrodes in the brain, the primary concern is to not harm or at least to minimize the risks with regards to the benefits. This risk-benefit approach is now adopted by ethics committees assessing the relevance of the protocols under consideration and assumes that it is possible to weight on a scale the various expected benefits, as well as the different damages, inconveniences or possible risks. Thus, the principle of the risk/benefit ratio consists in adding terms so radically heterogeneous between themselves as are pleasures and pains, converted to positive or negative units in order to anticipate the result of an action. While it appears as intuitively true, we should not forget that this is only an approximation. For invasive BCIs, although if opening a skull is a well-controlled surgical procedure, it still comprises significant risks (de Gray and Matta, 2010; Legnani et al., 2013; Kourbeti et al., 2015) and thus cannot be intuitively justified unless there is a real prospect for improvement of the patient's condition, which could not be achieved by other means.

3. Inf*ormed consent and patient's involvement*

It may be then tempting for the investigator of a clinical trial to limit descriptions by fear that no patient will agree to participate in a research protocol, which would have the consequence of slowing it down. Allowing patients to imagine that a purely cognitive trial would be beneficial to them and eventually therapeutic would therefore constitute a form of deceit, whatever the intention, maximizing the omnipotence of the one who knows but says nothing. It thus seems unethical to let patients fantasize about more promises than those the trial can deliver. The truth of the relationship between physicians and their patients comes at this price. Misleading someone is generally not ethical.

This trustful relationship is even more important that, in the case of invasive interfaces where patients are implanted for an indefinite period. Indeed, unlike short protocols in which patients are only involved transitorily, patients that are chronically implanted in the long term such as in (Hochberg et al., 2012) become become a major active participant of the research protocol, whose involvement – and often that of his family – is central to its success. As a result, patients become real contributors to the study, especially when they will not benefit directly from a BCI system that remains at a research and development stage.

[5] In France: Article R4127-109 of the public health code.
[6] Directive 2001/20/EC of the European Parliament and of the council of April 4, 2001.

4. Accessibility of BCIs

This issue is also linked to the problem of the financial cost of the commercial solutions for rehabilitation. In order to be available for people with disabilities whose financial resources are often limited because of a larger precarity[7] and a smaller salary on average[8], these technologies require a low-cost production on the long term. Especially for BCIs, for which some current prototypes are reaching very high costs, this implies that the research being undertaken should include the financial constraint allowing them to become accessible to the greatest number of people for whom they have been originally developed.

5. Modulating the brain activity with BCIs: what consequences?

Several studies suggest that the usage of a BCI system could lead to a modification of the neural substrate of the areas involved in the practiced tasks (Carmena et al., 2003; Ganguly and Carmena, 2009). Some studies on neurofeedback allow a subject to see in real time his brain activity – as in (Lachaux et al., 2007) –which is envisioned for teaching patients how to control their brain activity in order to correct their metanl, social or emotional behavior in a similar way to repeated physiotherapy sessions that allow correcting a motor behavior (Mercier-ganady et al., 2014). Beyond the clinical applications, such neurofeedback is envisioned for other purposes, such as improving performance. For instances, monitoring or strenghtening of vigilance has already been envisioned (Blankertz et al., 2010), as well as increasing the concentration on repetitive tasks such as for the security control of luggage in aiports (Müller et al., 2008). Depending on the objective sought-after, the ethical question of the usefulness of these practices, whose impact may concern people themselves or a third-party beneficiary, can then be raised (Vlek et al., 2012). In any case, it seems important that the subject remain free to choose whether to participate or not in these learning protocols.

6. Reliability and safety of BCIs

Given the quick advances in BCI technologies, one can assume that in a relatively near future, they will be parts of the daily lives of patients outside a protective clinical environment. User patients who want to be autonomous in their daily life, will thus need a BCI system both safe and reliable.

The system must be safe so as not to be disturbed accidentally or maliciously by the external environment. In addition, BCI systems transmit neural signals reflecting the thoughts and intents of action of their users, who may not wish to see these intimate and personal data being accessed by strangers. Thus, BCI systems ought to be protected, and this all the more that they will increasingly rely on wireless technologies (Guenther et al., 2009; Borton et al., 2013), where data can be more easily intercepted. In the event that this security problematics need to be seriously addressed, it is however not specific to BCIs only. Numerous objects increasingly more “intelligent” integrating embedded electronics and control software already occupy a

[7] 2011 data from French Ministry of Employment.
[8] Report n°45 of the agefiph in december 2013.

large part of our daily life including in a clinical context (for examples pacemakers or insulin pumps).

The system must also be reliable in order not to hurt its end-user or those that surround him or her, or even damage its environment. Indeed, neural signals may often be subject to fluctuations, which must be taken into account so as not to produce dangerous commands. This requires robust algorithms constraining the operation of the effector within predetermined limits, and thus limiting the actions that users can produce.

7. Responsibility when using BCIs

This issue is related to that of the responsibility that is involved when using a BCI system: Who should be held responsible for an accident caused by means of a BCI system? The user, the designer, or the distributor? This issue already gives rise to discussions that are all the more important that the spectrum of possible applications of the BCI is wide. There are uncertainties notably at multiple levels: The machine can incorrectly record or wrongly interpret the brain activity of the subject, but the brain activity of the subject can also be unconscious and not volitional. Subjects may also not fully control how their brain activity is interpreted and transformed into action (as a result of the algorithms being employed), which does not make them totally responsible for the actions that they would perform by using a BCI system. Thus the usual principle of analysis of the causal chain that consists of identifying whether the error comes from the user or from the machine seems actually harder to apply in the case of BCIs. Some authors consider that there is a responsibility gap where it is not possible to determine who is responsible (Lucivero and Tamburrini, 2008), while in opposition other authors point out that already existing laws and modes of reflection could be applied to the case of BCIs (Clausen, 2009; Haselager et al., 2009). Some authors porpose to implement certain rules before putting BCIs on the market, such as the obligation to measure and to assess the reliability of BCIs (Grübler, 2011). In all cases, the legal system should anticipate the emergence of BCIs in our daily lives so as to adapt itself accordingly.

IV. The human species

Brain-computer interfaces, whose main objective is the rehabilitation of functions in people suffering from severe disabilities, such as simply walking or talking again, remain largely uncommon nowadays. On the other hand, other types of interfaces between man and the machine multiply at high speed in our everyday environment: Touchscreens, speech or gesture recognition, facial recognition, augmented reality, etc. These non-invasive interfaces have been quickly democratized and integrated by society to the extent that they transform our daily life. Although the scientific limitations of BCI applications are still huge it is estimated that their first applications could appear on the market within a few decades (Blankertz et al., 2010; Nijboer et al., 2013). In this context, can we expect in a foreseeable future a strong influence of BCIs on the functioning of societies, or even on humanity itself?

1. BCIs as future means of enhancement?

Human beings have always sought to explore all the possibilities that were offered to them to increase their capabilities, for example by creating machines extending their abilities (moving quickly, flying, seeing at a distance, etc.), or by resorting to treatments extending their life. The boundary between therapy and enhancement is becoming increasingly blurred. In a world where the cult of performance is ever present, the choice of enhancement is attractive. In the pharmaceutical field for example, although the majority of drugs aim to heal, some substances are being used for the purpose of improving performance, such as physical performance, cognitive performance, sexual performance, etc. Surgery once reserved for the medical field is now used for aesthetics. The ever finer knowledge of neurobiological mechanisms also paves the way for the pharmacological improvement of performances (Farah, 2002). Other technologies are also used for enhancement: Some prostheses become today so efficient that they could allow people using them not only to circumvent their disabilities, but also to surpass and to reach physical performances enhanced and superior to the "norm" (Camporesi, 2008).

Because BCIs are still far from reproducing the natural performance of the human body, they have therefore not yet reached a stage allowing a person to be enhanced. However, some works suggest that this is potentially possible. For instance, a recent study shows how sensitivity to infrared light can be brought to rats through an implanted neuroprosthesis (Thomson et al., 2013), and other studies in humans suggest that BCI approaches can be used to improve certain cognitive performance such as attention (Gomez-Pilar et al., 2014) or short-term memorization (Burke et al., 2014).

If we consider for a moment the hypothesis that these technologies could indeed provide significant improvement of our faculties, this could impact the military field, where soldiers would improve their efficiency during their missions, but also all individuals eager to surpass themselves in their both private and professional lives. Even if this desire for enhancement for different purposes is not new, the potential of new technologies makes it possible to envisage a rupture of the level of enhancement that is offered. So far this level was maintained within the limitations of the body. The hybridization of the body with machines opens avenues toward a scaling of enhancement possibilities, which could result in a rethinking of the definition of the human.

2. The risk of transhumanism?

Such prospects feed the hopes of some of the transhumanist movements, which take the idea of human enhancement by technologies to the extreme, by advocating the fusion of man and machine to overcome not only disabilities, pain and diseases, but also any physical limitation of the body, such as aging and death, all being fatalities perceived as unnecessary and unwanted (Goffi, 2015). According to these movements, humans could find an extension of the human condition by merging with advanced technologies and thus controlling their own evolution toward a new augmented transhuman species. These perspectives are based on the fact that progress seems to follow an exponential development that could reach a critical point that some transhumanist theorists call "singularity" (Benderson, 2010). This progress, for some transhumanists, could now reach the "total prosthesis", enabling the instantaneous and

comprehensive understanding of human intentions by the machine, the establishment of transparent communication between two individuals communicating brain to brain without symbolic mediation, or even the complete transfer of the human spirit onto a machine (Lebedev et al., 2011; Neerdael, 2015). Although these perspectives sound very futuristic, they are however the focus of serious reflections and are largely financed.

The question may arise of whether these futuristic assumptions are not specifically intended to attract capital investors, themselves encouraged by the expectations of the general public (Neerdael, 2015).

3. Freedom and BCI

If the human species develops new technologies, it is to exploit and take advantage of them for its well-being. If the singularity principle proves to be true, mankind may become prisoner of the technologies that it will have developed for itself. Are we not already dependent on a large number of them, such as the means of communication by which we expect the other to be reactive without delay and at any time?

BCIs therefore require to consider the issue of freedom. Due to their functioning, these approaches consist of the real-time decoding of some of our intentions. While early works focused on motor intentions within a rehabilitation framework, the same methods are now striving, even if the goal is still far from being fully achieved, to decrypt more intimate information, such as perceived speech (Mesgarani et al., 2008), stored memories (Rissman et al., 2010) or even dreams (Horikawa et al., 2013). Despite these advances being significant and beneficial because they allow to better understand the functioning of our CNS, they provide the opportunity to explore specific information that people used to be free to keep for themselves. Will developed systems still be able to stop operating as soon as the user wants so? What usage, good or bad, will be made of the data collected about the user? The prospect that BCI technologies can easily be transposed beyond the field of research for social, economic (Ulman et al., 2015), political, military or even legal purposes (Wolpe et al., 2005), gives rise to the major ethical issue of their impact on the freedom of the individual or more generally of the human species. As suggested by Tennison et al. (Tennison and Moreno, 2012), if an enhancement technology may, for example, benefit to an employer, which could be, for example, a company looking for employees more focused on their work or an army aspiring for more effective soldiers (Kotchetkov et al., 2010), what freedom will then be left to an employee – faced with an employer or with the competition of those who have made the choice for enhancement – to comply to it or not? Subsequently, when a technology will have spread more widely to become a standard in everyday life, what freedom will an individual have, and more largely humans in general, to avoid it?

V. Conclusion

We are nowadays facing two major assumptions: either we consider the human being as existing within the limitations of a body, or we think the human being as a dynamics of undefined progress. In the first case, technological innovations must biologically comply with the limits of the human body, while in the second case, the body is only a support for technical innovations. These two assumptions allow distinguishing between the technological innovations that strive to serve more the purposes of the body from those that serve more the purposes of the will. However, the constraints that apply to technological innovations are not the same if these innovations aim at serving a body rather than a will. Serving a body requires respecting its functioning, its normativity, its limitations. Serving a will does not seem bound by the same constraints, the same limitations and the same internal normativity.

The goal of the ethical reflection presented here was not to tell which of these two assumptions is either true or false, but rather to try to accompany the current research in order to ask the important questions that could help enlighten minds. The issue of the welfare of the human species cannot oppose, in a probably too simplistic manner, bioconservatives and bioprogressists. The call for an international reflection group dedicated to these issues will certainly produce, as is the case for 60 years of global bioethics, some recommendations that will become laws governing these new technologies and their applications, but will not unfortunately always prevent that capital investors will sometimes go toward the best opportunities. Therefore, all the key players of these new technologies, including researchers, investors, vendors, and distributors, should probably prepare themselves to carefully examine the arising pro and contra arguments, without being able to rely on some allegedly higher ethical authority, since none of such is better than the one that their own conscience represents.

Part 6: Conclusions and Perspectives

In this thesis we considered speech restoration using a brain-computer interface (BCI). Indeed, brain-computer interfaces have shown promising results for restoring motor capabilities in paralysed patients. Similarly, they could be used to help aphasic patients suffering from severe paralysis to communicate, by allowing them to control a speech synthesizer from their brain activity. The goal of this thesis was thus to develop several aspects as proof of concept of this hypothesis. In this work we thus envisioned a speech BCI in which neural activity recorded in the speech motor cortex would be decoded into an intermediate articulatory representation of speech. This articulatory representation consisting in movements of the main articulators of the vocal tract could then be converted to speech using an articulatory-based speech synthesizer.

Main contributions and results

Toward this goal, we first built an articulatory-based speech synthesizer capable of synthesizing intelligible speech with few control parameters, then we investigate the cortical activity underlying speech production and the decoding of this activity into speech.

This was achieved by first recording a large dataset of synchronous acoustic and articulatory data. This corpus was recorded using electromagnetic articulography (EMA) in a French male speaker. EMA allowed to capture the trajectories of small sensors glued on the lips, jaw, tongue and soft palate of the speaker, with a spatial resolution inferior to the millimeter and a high temporal resolution. The final corpus consisted of more than 1,100 items, including all isolated vowels, vowel-consonant-vowel sequences and mostly full sentences, from short phonetically balanced sentences to long sentences extracted from newspapers. The articulatory trajectories were projected in the midsagittal plane of the speaker, resulting in a total of 14 time-varying articulatory features. The audio signals were also phonetically labeled at the phone level using a semi-automatic procedure. The whole dataset was publicly released during this thesis to facilitate the access to articulatory data to other research teams.

This articulatory-acoustic dataset was then used along with machine learning techniques to build a mapping that could convert articulatory trajectories into acoustic features, the so-called "articulatory-to-acoustic mapping". The acoustic features were obtained by computing mel-coefficients from the raw audio signal. In this thesis we chose to perform the articulatory-to-acoustic mapping using deep neural networks (DNNs). The training of DNNs caused several issues generally leading to poor solutions. Thus, we proposed a simple training method that proved to be efficient for the particular case of articulatory-to-acoustic mapping. In that approach, the training of the DNN was performed by successively adding layers to the network until its final configuration. This DNN training method, along with various other training methods and data pre-processing algorithms, was implemented in a custom software that has been used in various parts of this thesis work as well as in other projects. Once trained, the DNN could predict mel-coefficients from unpreviously seen articulatory data. These mel-coefficients could then be converted to speech using the MLSA filter. Results showed that intelligible speech could be synthesized with a word accuracy superior to 95%, and that the use of glottal activity could help to better discriminate consonants. Moreover, the synthesizer could

run fully in real-time, thus allowing its usage in applications such as a BCI for speech restoration.

We also compared our DNN-based mapping to a state-of-the-art approach using gaussian mixture regression. In particular, we evaluated how the number of input parameters and their noise level impacted the synthesis. Indeed, BCIs can typically decode about ten degrees of freedom with relatively good accuracy. A speech synthesizer controlled by BCI should thus have few parameters and be robust to fluctuations of these parameters. This was simulated by performing dimensionality reduction on the input articulatory data and by adding artificial noise to these inputs. Results showed that the DNN-based mapping was more robust to noisy inputs than the GMM-based mapping. Moreover, the results also showed that intelligible speech could still be achieved with about ten articulatory parameters.

This synthesizer was then controlled in real-time from the articulatory movements of several subjects, in a closed-loop paradigm. Indeed, it remained unknown whether a given articulatory-based speech synthesizer built from articulatory-acoustic data obtained in one particular reference speaker could be controlled in real time by any other speaker to produce intelligible speech. We thus had the EMA data of several speakers recorded in real-time in order to control the articulatory-based synthesizer, while they were given the synthesis feedback through earphones. We proposed a simple calibration method allowing to compensate for the differences between the reference speakers and the new speakers. This method consisted in synchronously articulating a small set of short sentences extracted from the original articulatory-acoustic dataset. Thus, the reference articulatory movements for these particular sentences were known, which allowed to compute a linear mapping between the new speakers' articulatory space and the reference speaker's articulatory space, the so called "articulatory-to-articulatory mapping". The results showed that the speech synthesizer could be controlled in real-time by the different speakers to produce not only vowels, but also intelligible consonants and some sentences. This result was achieved despite the fact that the articulatory-to-articulatory mapping was erroneous, further confirming the robustness of the proposed synthesis method. The overall intelligibility of the synthesis was however limited when compared to the offline condition, especially because of a poor synthesis of plosive consonants.

We then considered the mapping of speech-related brain areas in patients undergoing awake brain surgery at the university hospital of Grenoble. Indeed, one treatment for brain tumors consists in removing the tumor areas and cells while preserving the sensory-motor functions of the patient. Mapping the functional areas of speech during surgery could help identify the areas to be preserved. Moreover, the localization of speech-related brain areas can help optimize the positioning of micro-electrode arrays used in a BCI for speech rehabilitation. This was as well a good opportunity to record neural data during speech production and imagination. We thus developed an approach to map speech-related brain activity, directly during awake brain surgery, and on the neurosurgeon's view of the operative field. First, automatic speech detection was performed using the signal from a microphone placed next to the patient in order to identify speech production data segments. This allowed to compute the mean spectral power for both speech production and rest periods using short-time Fourier transform. Speech-related brain areas were then identified by quantifying differences in spectral power between speech and silence segments, the so called speech-silence-ratio. Statistical tests were used to assess the significance of these differences in order to only map the relevant changes in neural activity. The speech-silence-ratio was then mapped directly on a picture of

the neurosurgeon's view of the operative field using spline interpolation. This first required to coregister the electrodes on the anatomy. This was achieved by first identifying some of the electrodes on a picture of the operative field with the electrodes visible. The positions of the other electrodes were then approximated using spline interpolation. Then, anatomical landmarks were identified and the picture with and the picture without the electrodes visible, allowing to transpose the electrodes positions to the picture where they were not visible. We could thus map speech-related brain activity on a picture of the operative field in which the electrodes were not present. We also proposed a simple method to estimate the locations of the electrodes on a reconstructred cortical surface from MRI data. All these methods were implemented into a software, called ClientMap, with the help of two interns that I supervised. This software was then tested in two patients undergoing awake brain surgery for a tumor removal, which tumor was located next to the speech motor cortex. The developped software was fully functional and allowed to map the speech-related brain activities of the patients. Results were consistent between both patients and exhibited a beta desynchronization as well as an increase in low and high gamma bands in the speech motor cortex during speech production. This was also observed during imagination of speech and also during listening periods.

We then started to investigate how speech could be decoded from neural activity. In a first step, we considered the decoding of speech intention, i.e. predicting if the patient produces or intent to produce speech. Results showed that speech intention could be predicted with accuracy up to 94%. Moreover, a model only trained on overt speech data was also able to predict covert (imagined) speech periods with an accuracy of 78%. This suggests that overt and covert speech at least share a partially common neural representation in the speech motor cortex. In a second step we addressed the decoding of voicing activity, i.e. predicting if the vocal folds are vibrating or not during speech production. Results were still above chance level but lower than when decoding speech intention (about 75% accuracy). In a third step, we aimed at decoding articulatory trajectories from the neural data. Here the articulatory trajectories consisted of 14 time-varying coordinates obtained from electromagnetic articulography. Since the articulatory trajectories of the patient were not recorded during surgery, we proposed an approach to estimate them. This approach relied on the fact that the patient produced sentences for which articulatory trajectories and the audio of another subject were previously recorded and known. We used dynamic time warping to align the audio of this subject with the patient's audio. The resulting time-warping was then applied to the known articulatory trajectories in order to align them with the neural data of the patient. Results showed that several articulatory trajectories could be decoded above chance level, in particular the vertical position of the jaw, the vertical position of the tongue tip, the horizontal position of the back of the tongue and the horizontal position of the velum. Moreover, the results were best when no delay between the neural and the articulatory data was added, and when considering at least 100ms chunks of neural data. Reducing the dimensionality of the neural feature space also helped improve the decoding accuracy. These decoded articulatory trajectories were then converted to acoustic features using the articulatory-based speech synthesis. However, the decoding accuracy remained too limited to generate intelligible speech. To test the hypothesis that decoding articulatory trajectories as input to an articulatory-based speech synthesizer could be more efficient that directly predicting acoustic parameters to synthesize speech, we as well considered the decoding of acoustic features from neural activity. Results showed that several acoustic features could be decoded above chance level, in particular the first, second, fifth and seventh mel coefficients. As for the

decoding of articulatory trajectories, results suggested that best accuracy could be achieved without any delay between the neural and acoustic data, when considering neural data chunks of at least 100ms, and when using dimensionality reduction techniques. We then compared the acoustic features depending on whether they were decoded directly from neural data or from articulatory trajectories estimated from neural data. Results showed that most of the acoustic features were best decoded when passing by the intermediate articulatory representation, which supports our main hypothesis.

Finally, we addressed several ethical issues arising with the usage and development of brain-computer interfaces (BCIs). We considered three levels of ethical questionings, dealing respectively with the animal, the human being, and the human species. For the animal, we addressed the common issue of pain and suffering when performing animal experiments, but also that of the control of animals by BCIs. For the human being, we discussed the involvement of the patients in clinical research along with the hopes that it can arouse. We also addressed the case of the future users of BCIs, and several aspects regarding their access to these technologies, their reliability and safety, but also the impacts of their usage on the user himself and its surrounding environment. Finally, in a perspective of generalized use of BCIs, we discussed the impact it could have on the human species, especially for non-clinical usages of BCIs. The goal of this reflection was not to provide answers but rather to raise ethical questions to bear in mind while conducting research on BCIs.

Perspectives

This thesis work presents several main steps toward a brain-computer interface for speech rehabilitation using articulatory data: the development of an articulatory-based speech synthesizer, the identification of the speech-related cortical areas and the decoding or articulatory trajectories from neural data. However, the proposed approaches here should be improved in several aspects.

First, the offline articulatory-based speech synthesis might be improved in several ways. Indeed, while intelligible speech could be synthesized with an open-vocabulary word accuracy above 90%, the synthesis accuracy on consonants was only about 70%. In particular, plosive consonants had the lowest recognition accuracy, and future works should essentially focus on improving their synthesis quality. This could be achieved by efficiently detecting specific constrictions from the articulatory data, which would then serve as additional inputs to the synthesizer. The overall synthesis intelligibility could also be improved by using a priori knowledge, such as a dictionary of all the words of the target language, combined with grammatical rules or word sequences probabilities. Such a priori knowledge could improve the synthesis by constraining it to a specific language or limited vocabulary. This could be done for instance by combining the advantage of deep neural networks for regression of continuous variables, with the advantages of hidden markov models for modeling sequences of discretes states, such as phones, words or semantic units.

Similar conclusions also emerged from the results of the real-time control of this synthesizer from articulatory movements of new subjects. Indeed, most synthesis errors were similar to that of the offline synthesis, indicating that using the synthesizer in a closed-loop paradigm mainly emphasized the already existing confusions. However, the presence of

additional minor confusions suggests that other aspects might also be improved, such as the calibration approach proposed to compensate for the articulatory differences between subjects. Subjects may also adapt differently to the articulatory-to-articulatory mapping errors and find behavioral strategies to compensate for these errors. Here, the different subjects had about half an hour training time to improve their control of the synthesizer but with no significant results. One subject however reported self-improvement. Thus, future studies should further explore the possibilities of improvement through longer training periods. This would require to find a way to record stable EMA signals between sessions, for instance using unsupervised calibration methods.

In this thesis, we also presented a method to localize and visualize speech-related brain activity in real-time, during awake brain surgery. While being efficient, this method could be improved in several points. In particular, the automatic speech detection could be improved using statistical models of speech and silence audio signals. The mapping workflow might also be improved by automatizing or semi-automizing the coregistration of the electrodes on the anatomy and the reconstructed cortical surface. This could be achieved by using computer vision techniques and by taking into account the geometry of the cortical surface when localizing the electrodes. Moreover, results from the mapping of speech-related brain activity suggested that there is at least a partially common neural representation of overt speech, covert speech and speech listening in the speech motor cortex. Future experiments should confirm this hypothesis, in particular by studying covert speech with denser recordings. Further experiments should also consider a finer mapping of speech-related activity, for instance at the phone level or according to the place of articulation. This could be achieved by performing automatic speech recognition to directly segment speech and map the neural activity for each type of speech unit.

Finally, we presented three different decoding steps required to build a BCI for speech rehabilitation based on articulatory data: detecting speech intention, decoding voicing activity, and decoding articulatory trajectories. While overt and covert speech intentions were correctly detected, results showed that listening periods were also detected as speech intention. Here, this was observed in one patient that only had four relatively-large electrodes recording the neural activity. Thus, future work should further study this aspect and in particular investigate the differences and similarities of the neural activity in the speech motor cortex during speech production and speech listening. Decoding the voicing activity could also be performed with above chance level. However, the decoding accuracy remained limited as compared to that of the speech intention detection. This might be due to the limited number of electrodes actually covering the speech motor cortex, and should be further investigated in future experiments. Future work should also consider the decoding of the voicing activity during covert speech, since it was only performed for overt speech in this work. Finally, we showed that some articulatory trajectories ad acoustic features could be decoded above chance level. However, this did not result in intelligible speech synthesis. While these results are encouraging, future work should focus on improving the decoding accuracy. This could be achieved by recording larger datasets in order to use more advanced machine learning techniques, and by using denser recordings at the cellular level, for instance micro-electrode arrays. Moreover, we compared the indirect decoding of speech using articulatory trajectories to a direct decoding of acoustic features from neural data recorded in the speech motor cortex. The results suggested that best decoding accuracy was achieved using an intermediate articulatory representation of speech,

which supports our main hypothesis. However, this was only achieved in one patient, and the decoding accuracy remained limited. This result must thus be confirmed by future experiments. If this hypothesis is proven right, the three different decoding stages presented in this work could be combined in a BCI for speech rehabilitation, as schematized in **Fig. 94**.

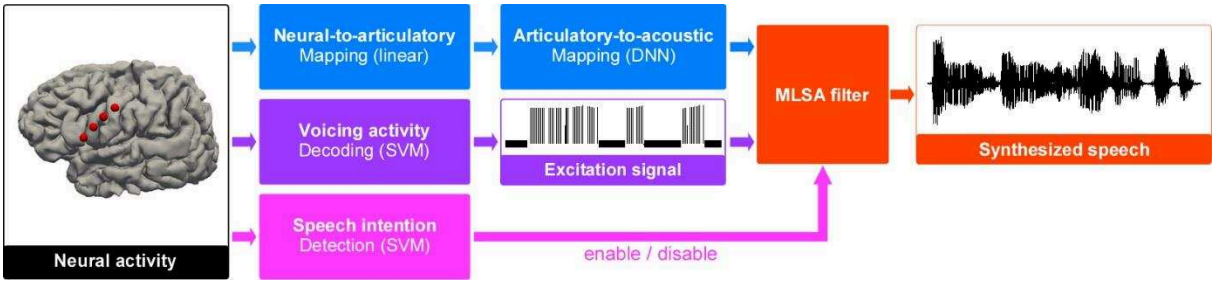


Fig. 94: Conceptual view of articulatory-based speech synthesis from neural data. First, the neural activity is used to detect speech intention in order to enable or disable the speech synthesis (pink path). If speech intention is detected, voicing activity is decoded from the neural data to generate an appropriate excitation signal for voiced and unvoiced sounds (purple path). This neural activity is then decoded into articulatory trajectories which are then converted to mel coefficients using the articulatory-based speech synthesizer (blue path). Finally, the MLSA filter combine the mel coefficients and the excitation signal to synthesize speech (orange path).

Annexes

Annex 1: List of sentences for the evaluation of the reference offline synthesis

Francfort a reculé de un virgule quarante et un pourcent.
Au moins une sévère leçon.
Les préparatifs vont bon train.
Comme un regain en somme.
Allègre portrait d'un tueur psychopathe.
Les prix flambent dans l'hôtellerie.
Vous êtes un homme d'appareil.
Les travaux des bûcherons.
Le reste sera européen.
Le trafic d'être humain augmente.
Leur avenir semble aussi incertain.
Encore moins à son affiche.
Tel est le message implicite.
Et la clôt fort dignement.
Un homme toujours de dos.
Elle se trompe de débat.
Il y en a un.
Avec grandeur et densité.
Voilà où nous en sommes.
Elle en épaissit les outrances.
Les besoins y sont énormes.
Le quinze de France.
Ne pas rater le coach.
Le dollar chute, l'euro s'envole.
J'en suis totalement incapable.
Les autres réorganisations devraient suivre.
Les gammes familiales s'étoffent.
L'œuf du serpent aurait éclos.
Jugement le dix-neuf janvier.
Il lui faudra huit ans.

Annex 2: List of sentences from the spontaneous conversation during the real-time control of the synthesizer

Subject	Original sentence pronounced by the subject
Subject 1	Voilà j'ai fini, qu'en penses-tu de cette période ?
Subject 1	Bienvenue à Gipsa-lab et à Clinatec.
Subject 1	J'ai l'impression que ça marche mieux.
Subject 1	Je vais commencer l'entrainement numéro deux.
Subject 1	C'est vraiment beaucoup mieux que la dernière fois.
Subject 1	Fin du troisième entrainement.
Subject 1	J'aime bien le chocolat.
Subject 1	BrainSpeak est un joli projet.
Subject 1	La chartreuse est un alcool fabriqué en Isère.
Subject 1	Bravo, super, c'est vraiment génial.
Subject 1	J'ai l'impression que je ne progresse plus.
Subject 1	Je fais des phrases maintenant.
Subject 1	Quel est ton plat préferré ?
Subject 1	As-tu été au cinéma ce weekend ?
Subject 1	Aimes-tu les cuisses de grenouilles ?
Subject 1	Oui c'est bien cela.
Subject 2	Je vais être papa, je suis très content, c'est une bonne occasion de vous l'annoncer.
Subject 2	J'ai un champignon sur mon ongle droit.
Subject 2	J'espère qu'on va écrire un super papier avec cette manip qui déchire.
Subject 2	Ceci est une interface de communication en parole silencieuse.

Bibliography

Abdoun O, Joucla S, Mazzocco C, Yvert B (2011) NeuroMap: A Spline-Based Interactive Open-Source Software for Spatiotemporal Mapping of 2D and 3D MEA Data. Front Neuroinform 4:119.

Aristote (n.d.) Ethique à Nicomaque.

Atal BS (2006) The history of linear prediction. IEEE Signal Process Mag 23:154–158.

Aziz-Zadeh L, Cattaneo L, Rochat M, Rizzolatti G (2005) Covert speech arrest induced by rTMS over both motor and nonmotor left hemisphere frontal sites. J Cogn Neurosci 17:928–938.

Badin P, Gabioud B, Beautemps D, Lallouache TM, Bailly G, Maeda S, Zerling JP, Brock G (1995) Cineradiography of VCV sequences : Articulatory-acoustic data for a speech production model.

Baer T, Alfonso PJ, Honda K (1988) Electromyographie of the tongue muscles during vowels. Annu Bull Res Inst Logop Phoniatr 22:7–19.

Basho S, Palmer ED, Rubio MA, Wulfeck B, M??ller RA (2007) Effects of generation mode in fMRI adaptations of semantic fluency: Paced production and overt speech. Neuropsychologia 45:1697–1706.

Basirat A, Sato M, Schwartz JL, Kahane P, Lachaux JP (2008) Parieto-frontal gamma band activity during the perceptual emergence of speech forms. Neuroimage 42:404–413.

Bay H, Tuytelaars T, Van Gool L (2006) SURF: Speeded up robust features. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 3951 LNCS:404–417.

Baykara E, Ruf CA, Fioravanti C, Käthner I, Simon N, Kleih SC, K??bler A, Halder S (2016) Effects of training and motivation on auditory P300 brain-computer interface performance. Clin Neurophysiol 127:379–387.

Beautemps D, Badin P, Bailly G (2001) Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. J Acoust Soc Am 109:2165–2180 Available at: http://scitation.aip.org/content/asa/journal/jasa/109/5/10.1121/1.1361090 [Accessed March 18, 2014].

Bechet F (2001) LIA_PHON - Un systeme complet de phonetisation de textes. Trait Autom des Langues 42.

Bekoff M (2000) Animal Emotions: Exploring Passionate Natures. Bioscience 50:861 Available at: http://bioscience.oxfordjournals.org/cgi/doi/10.1641/0006-3568(2000)050[0861:AEEPN]2.0.CO;2 [Accessed November 30, 2016].

Ben Youssef A, Badin P, Bailly G, Heracleous P (2009) Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models. Proc Interspeech:2255–2258.

Benabid AL, Pollak P, Hoffmann D, Gervason C, Hommel M, Perret JE, de Rougemont J, Gao DM (1991) Long-term suppression of tremor by chronic stimulation of the ventral intermediate

thalamic nucleus. Lancet 337:403–406 Available at: http://www.thelancet.com/article/014067369191175T/fulltext [Accessed November 30, 2016].

Benderson B (2010) Transhumain, Payot. Paris.

Bengio Y, Lamblin P, Popovici D, Larochelle H (n.d.) Greedy Layer-Wise Training of Deep Networks.

Birkholz P, Jackel D, Kroger KJ (2006) Construction And Control Of A Three-Dimensional Vocal Tract Model. 2006 IEEE Int Conf Acoust Speech Signal Process Proc 1:873–876.

Bishop CM, Christopher M. B (2006) Pattern Recognition and Machine Learning. New Yor: Springer-Verlag.

Black a. W, Zen H, Tokuda K (2007) Statistical Parametric Speech Synthesis. 2007 IEEE Int Conf Acoust Speech Signal Process - ICASSP '07 4.

Blankertz B, Tangermann M, Vidaurre C, Fazli S, Sannelli C, Haufe S, Maeder C, Ramsey L, Sturm I, Curio G, Müller K-R (2010) The Berlin Brain-Computer Interface: Non-Medical Uses of BCI Technology. Front Neurosci 4:198 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3002462&tool=pmcentrez&rendertype=abstract [Accessed November 30, 2014].

Bocquelet F, Hueber T, Girin L, Badin P, Yvert B (2014) Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), pp 2288–2292.

Bocquelet F, Hueber T, Girin L, Savariaux C, Yvert B (2015) Real-time control of a dnn-based articulatory synthesizer for silent speech conversion: A pilot study. Proc Annu Conf Int Speech Commun Assoc INTERSPEECH 2015-Janua:2405–2409.

Bocquelet F, Hueber T, Girin L, Savariaux C, Yvert B (2016a) Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces. PLoS Comput Biol 12.

Bocquelet F, Hueber T, Girin L, Savariaux C, Yvert B (2016b) BY2014 articulatory-acoustic dataset. Zenodo:1–28 Available at: https://zenodo.org/record/154083 [Accessed October 11, 2016].

Bocquelet F, Piret G, Aumonier N, Yvert B (2016c) Ethical Reflections on Brain-Computer Interfaces. In: Brain-Computer Interfaces 2, pp 259–288. Hoboken, NJ, USA: John Wiley & Sons, Inc. Available at: http://doi.wiley.com/10.1002/9781119332428.ch15 [Accessed March 1, 2017].

Bonferroni (1936) Teoria statistica delle classi e calcolo delle probabilità. Pubbl del R Ist Super di Sci Econ e Commer di Firenze 8.

Bonnet L, Lotte F, Anatole L, Rennes I, Bordeaux I (2013) Two Brains , One Game : Design and Evaluation of a Multi-User BCI Video Game Based on Motor Imagery To cite this version : Two Brains , One Game : Design and Evaluation of a Multi-User BCI Video Game Based on Motor Imagery. 2:1–13.

Bonnet S et al. (2012) NeuroPXI: A real-time multi-electrode array system for recording, processing and stimulation of neural networks and the control of high-resolution neural implants for rehabilitation. Irbm 33:55–60 Available at: http://dx.doi.org/10.1016/j.irbm.2012.01.013.

Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E (2014) Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J Neurosci 34:4548–4557.

Borton DA, Yin M, Aceros J, Nurmikko A (2013) An implantable wireless neural interface for recording cortical circuit dynamics in moving primates. J Neural Eng 10:026010 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23428937 [Accessed November 30, 2016].

Bouchard KE, Conant DF, Anumanchipalli GK, Dichter B, Chaisanguanthum KS, Johnson K, Chang EF (2016) High-Resolution, Non-Invasive Imaging of Upper Vocal Tract Articulators Compatible with Human Brain Recordings. PLoS One 11:e0151327 Available at: http://dx.plos.org/10.1371/journal.pone.0151327.

Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation. Nature 495:327–332 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23426266 [Accessed September 19, 2013].

Boussen S, Velly L, Benar C, Metellus P, Bruder N, Tr??buchon A (2016) In Vivo Tumour Mapping Using Electrocorticography Alterations During Awake Brain Surgery: A Pilot Study. Brain Topogr 29:766–782.

Brandmeyer A, Farquhar JDR, McQueen JM, Desain PWM (2013) Decoding speech perception by native and non-native speakers using single-trial electrophysiological data. PLoS One 8:e68261.

Brown S, Laird AR, Pfordresher PQ, Thelen SM, Turkeltaub P, Liotti M (2009) The somatotopy of speech: Phonation and articulation in the human motor cortex. Brain Cogn 70:31–41 Available at: http://dx.doi.org/10.1016/j.bandc.2008.12.006.

Brown S, Ngan E, Liotti M (2008) A larynx area in the human motor cortex. Cereb Cortex 18:837–845 Available at: http://www.ncbi.nlm.nih.gov/pubmed/17652461 [Accessed October 21, 2013].

Brugnara F, Falavigna D, Omologo M (1993) Automatic segmentation and labeling of speech based on Hidden Markov Models. Speech Commun 12:357–370.

Brumberg JS, Wright EJ, Andreasen DS, Guenther FH, Kennedy PR (2011) Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. Front Neurosci 5:65 Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21629876&retmode=ref&cmd=prlinks\npapers2://publication/doi/10.3389/fnins.2011.00065.

Brunner P, Joshi S, Briskin S, Wolpaw JR, Bischof H, Schalk G (2010) Does the “P300” speller depend on eye gaze? J Neural Eng 7:056013 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20858924 [Accessed November 30, 2016].

Burke JF, Merkow MB, Jacobs J, Kahana MJ, Zaghloul KA (2014) Brain computer interface to enhance episodic memory in human participants. Front Hum Neurosci 8:1055 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25653605 [Accessed November 30, 2016].

Buzsáki G, Draguhn A (2004) Neuronal oscillations in cortical networks. Science 304:1926–1929 Available at: http://www.ncbi.nlm.nih.gov/pubmed/15218136.

Camporesi S (2008) Oscar Pistorius, enhancement and post-humans. J Med Ethics 34:639 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18757629 [Accessed November 30, 2016].

Canolty RT, Soltani M, Dalal SS, Edwards E, Dronkers NF, Nagarajan SS, Kirsch HE, Barbaro NM, Knight RT (2007) Spatiotemporal dynamics of word processing in the human brain. Front Neurosci 1:185–196.

Canto-Sperber M, Ogien R (2004) La Philosophie morale, Presses un. Paris.

Carmena JM, Lebedev M a, Crist RE, O'Doherty JE, Santucci DM, Dimitrov DF, Patil PG, Henriquez CS, Nicolelis M a L (2003) Learning to control a brain-machine interface for reaching and grasping by primates. PLoS Biol 1:E42 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=261882&tool=pmcentrez&rendertype=abstract [Accessed October 18, 2013].

Castermans T, Duvinage M, Cheron G, Dutoit T (2014) Towards Effective Non-Invasive Brain-Computer Interfaces Dedicated to Gait Rehabilitation Systems. Brain Sci 4:1–48 Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4066236/.

Chan TF, Golub GH, LeVeque RJ (1983) Algorithms for Computing the Sample Variance: Analysis and Recommendations. Am Stat 37:242.

Chang Y-W, Hsieh C-J, Chang K-W, Ringgaard M, Lin C-J (2010) Training and Testing Low-degree Polynomial Data Mappings via Linear SVM. J Mach Learn Res 11:1471–1490 Available at: http://www.csie.ntu.edu.tw/~cjlin/papers/lowpoly_journal.pdf.

Charpentier FJ, Stella MG (1986) Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: ICASSP, pp 2015–2018 Available at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1168657&tag=1.

Cheng C, Huo X, Ghovanloo M (2009) Towards A Magnetic Localization System for 3-D Tracking of Tongue Movements in Speech-Language Therapy. In: Conf Proc IEEE Eng Med Biol Soc., pp 563–566.

Cheung C, Hamiton LS, Johnson K, Chang EF (2016) The auditory representation of speech sounds in human motor cortex. Elife 5:1–19.

Choupan J, Hocking J, Johnson K, Reutens D, Yang Z (2013) Brain Decoding Based on Functional Magnetic Resonance Imaging Using Machine Learning: A Comparative Study. Int J Mach Learn Comput 3:132–136 Available at: http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=35&id=285.

Clausen J (2009) Man, machine and in between. Nature 457:1080–1081 Available at: http://www.nature.com/doifinder/10.1038/4571080a [Accessed November 30, 2016].

Cler MJ, Nieto-Castanon a, Guenther FH, Stepp CE (2014) Surface electromyographic control of speech synthesis. Eng Med Biol Soc (EMBC), 2014 36th Annu Int Conf IEEE:5848–5851.

Cogan GB, Thesen T, Carlson C, Doyle W, Devinsky O, Pesaran B (2014) Sensory-motor transformations for speech occur bilaterally. Nature:0–6 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24429520 [Accessed January 20, 2014].

Cohen LB (1973) Changes in Neuron Potential Propagation Structure During Action and Synaptic Transmission. Physiol Rev 53:373–418.

Collinger JL, Wodlinger B, Downey JE, Wang W, Tyler-Kabara EC, Weber DJ, McMorland AJC, Velliste M, Boninger ML, Schwartz AB (2013) High-performance neuroprosthetic control by an individual with tetraplegia. Lancet 381:557–564 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23253623 [Accessed September 23, 2013].

Congedo M, Goyat M, Tarrin N, Varnet. L, Rivet B, Ionescu G, Jrad N, Phlypo R, Acquadro M, Jutten C (2011) "Brain Invaders": a prototype of an open-source P300-based video game working with the OpenViBE platform. 5th Int BCI Conf Graz, Austria, 280-283 2011:1–6 Available at: http://hal.archives-ouvertes.fr/hal-00641412/.

Correia J, Formisano E, Valente G, Hausfeld L, Jansma B, Bonte M (2014) Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. J Neurosci 34:332–338.

Correia JM, Jansma BMB, Bonte M (2015) Decoding Articulatory Features from fMRI Responses in Dorsal Speech Regions. J Neurosci 35:15015–15025.

Coyle SM, Ward TE, Markham CM (2007) Brain–computer interface using a simplified functional near-infrared spectroscopy system. J Neural Eng 4:219–226 Available at: http://www.ncbi.nlm.nih.gov/pubmed/17873424 [Accessed December 6, 2016].

Cramer G (1750) Introduction à l'Analyse des lignes Courbes algébriques, Geneva: Eu.

Cui X, Bray S, Bryant D, Glover G, Reiss A (2011) A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. Neuroimage 54:2808–2821.

Dang J, Honda K (2004) Construction and control of a physiological articulatory model. J Acoust Soc Am 115:853–870.

David OE, Aviv T, Greental I (2014) Genetic Algorithms for Evolving Deep Neural Networks. :1451–1452.

De Gray LC, Matta BF (2010) Acute and chronic pain following craniotomy. Curr Opin Anaesthesiol 23:551–557 Available at: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L51037545\nhttp://dx.doi.org/10.1097/ACO.0b013e32833e15b9\nhttp://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=09527907&id=doi:10.1097/ACO.0b013e32833e15b9&atitle=Acute+and+chronic+pain.

De Mareüil P, Rilliard a, Allauzen a (2008) A diachronic study of prosody through French audio archives. Speech Prosody:531–534.

Denby B, Oussar Y, Dreyfus G, Stone M (2006) Prospects for a Silent Speech Interface using Ultrasound Imaging. Proc 2006 IEEE Int Conf Acoust Speed Signal Process Proc 1:I – 365 – I – 368 Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1660033.

Denby B, Schultz T, Honda K, Hueber T, Gilbert JM, Brumberg JS (2010) Silent speech interfaces. Speech Commun 52:270–287 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167639309001307.

Denby B, Stone M (2004) Speech synthesis from real time ultrasound images of the tongue. In: IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP04. Montréal.

Deng J, Newton NM, Hall-craggs MA, Shirley RA, Linney AD, Lees WR, Rodeck CH, Mcgrouther DA (2000) Early report Novel technique for three-dimensional visualisation and quantification of deformable , moving soft-tissue body parts. 356:127–131.

Deng S, Srinivasan R, Lappas T, D'Zmura M (2010) EEG classification of imagined syllable rhythm using Hilbert spectrum methods. J Neural Eng 7:046006 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20551510 [Accessed December 7, 2016].

Diener L, Herff C, Janke M, Schultz T (2016) An Initial Investigation into the Real-Time Conversion of Facial Surface {EMG} Signals to Audible Speech. 38th Int Conf IEEE Eng Med Biol Soc:888–891.

Dietrich D, Lang R, Bruckner D, Fodor G, Müller B (2010) Limitations, possibilities and implications of Brain-Computer Interfaces. 3rd Int Conf Hum Syst Interact HSI'2010 - Conf Proc:722–726.

Dudley HW (1939) Signal transmission. Available at: http://www.google.com/patents/US2151091?hl=fr [Accessed March 21, 1939].

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

Duffau H, Leroy M, Gatignol P (2008) Cortico-subcortical organization of language networks in the right hemisphere: An electrostimulation study in left-handers. Neuropsychologia 46:3197–3209.

Duffau H, Moritz-Gasser S, Mandonnet E (2014) A re-examination of neural basis of language processing: Proposal of a dynamic hodotopical model from data provided by brain stimulation mapping during picture naming. Brain Lang 131:1–10 Available at: http://dx.doi.org/10.1016/j.bandl.2013.05.011.

Engel AK, Moll CKE, Fried I, Ojemann G a (2005) Invasive recordings from the human brain: clinical insights and beyond. Nat Rev Neurosci 6:35–47 Available at: http://www.ncbi.nlm.nih.gov/pubmed/15611725 [Accessed August 8, 2014].

Engelhard B, Ozeri N, Israel Z, Bergman H, Vaadia E (2013) Inducing Gamma Oscillations and Precise Spike Synchrony by Operant Conditioning via Brain-Machine Interface. Neuron 77:361–375.

Engineer CT, Perez CA, Chen YH, Carraway RS, Reed AC, Shetake JA, Jakkamsetti V, Chang KQ, Kilgard MP (2008) Cortical activity patterns predict speech discrimination ability. Nat Neurosci 11:603–608.

Engwall O (1999) Modelling of the vocal tract in three dimensions. Proc Eurospeech 1:113–116.

Engwall O (2003) A revisit to the application of MRI to the analysis of speech production - Testing our assumptions. In: Proc. 6th Int. Seminar on Speech Production (ISSP), pp 43–48. Sydney, Australia.

Fabre D, Hueber T, Bocquelet F, Badin P (2015) Tongue Tracking in Ultrasound Images using EigenTongue Decomposition and Artificial Neural Networks. In: Proceedings of the Interspeech Conference.

Fagan MJ, Ell SR, Gilbert JM, Sarrazin E, Chapman PM (2008) Development of a (silent) speech recognition system for patients following laryngectomy. Med Eng Phys 30:419–425.

Fant G (1975) Vocal-tract area and length perturbations. Stl-Qpsr Available at: http://www.speech.kth.se/prod/publications/files/qpsr/1975/1975_16_4_001-014.pdf.

Farah MJ (2002) Emerging ethical issues in neuroscience. Nat Neurosci 5:1123–1129.

Farwell L a, Donchin E (1988) Talking Off the Top of Your Head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr Clin Neurophysiol 70:510–523.

Fenster A (2001) Three-dimensional ultrasound imaging. Phys Med Biol 46:R67–R99.

Fetz EE (1969) Operant conditioning of cortical unit activity. Science (80-) 163:955–958.

Fetz EE (2007) Volitional control of neural activity : implications for brain – computer interfaces. Society 579:571–579 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2151376&tool=pmcentrez&rendertype=abstract.

Fisher CG (1968) Confusions among visually perceived consonants. J Speech Hear Res 11:796–804 Available at: http://www.ncbi.nlm.nih.gov/pubmed/5719234 [Accessed June 30, 2015].

Flanagan J, Ishizaka K, Shipley K (1975) Synthesis of speech from a dynamic model of the vocal cords and vocal tract. Bell Syst Tech J 54:485–506 Available at: http://onlinelibrary.wiley.com/doi/10.1002/j.1538-7305.1975.tb02852.x/abstract.

Fontecave Jallon J, Berthommier F (2009) A semi-automatic method for extracting vocal tract movements from X-ray films. Speech Commun 51:97–115 Available at: http://linkinghub.elsevier.com/retrieve/pii/S016763930800099X [Accessed June 25, 2015].

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

Formisano E, Martino F De, Bonte M, Goebel R (2008) "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. Science (80-) 322:970–973.

Friederici AD (2011) The brain basis of language processing: from structure to function. Physiol Rev 91:1357–1392.

Gales M, Young S (2007) The Application of Hidden Markov Models in Speech Recognition. Found Trends® Signal Process 1:195–304 Available at: http://www.nowpublishers.com/product.aspx?product=SIG&doi=2000000004 [Accessed March 20, 2014].

Ganguly K, Carmena JM (2009) Emergence of a stable cortical map for neuroprosthetic control. PLoS Biol 7:e1000153 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2702684&tool=pmcentrez&rendertype=abstract [Accessed October 9, 2013].

Gavin HP (2013) The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems. Dep Civ Environ Eng Duke Univ:1–17.

Geranmayeh F, Wise RJS, Mehta A, Leech R (2014) Overlapping Networks Engaged during Spoken Language Production and Its Cognitive Control. J Neurosci 34:8728–8740.

Gilbert JM, Rybchenko SI, Hofe R, Ell SR, Fagan MJ, Moore RK, Green P (2010) Isolated word recognition of silent speech using magnetic implants and sensors. Med Eng Phys 32:1189–1197 Available at: http://dx.doi.org/10.1016/j.medengphy.2010.08.011.

Giraud A-L, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511–517 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22426255 [Accessed April 28, 2014].

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. 9:249–256.

Goffi J-Y (2015) Transhumain. In: Encyclopédie du trans/poshthumanisme, L'humain et ses préfixes, Vrin. (Hottois G, Missa J-N, Perdal L, eds), pp 156–163. Paris.

Gold C, Henze D a, Koch C, Buzsáki G (2006) On the origin of the extracellular action potential waveform: A modeling study. J Neurophysiol 95:3113–3128 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16467426 [Accessed October 31, 2014].

Gomez-Pilar J, Corralejo R, Nicolas-Alonso LF, Álvarez D, Hornero R (2014) Assessment of neurofeedback training by means of motor imagery based-BCI for cognitive rehabilitation. Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2014:3630–3633 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25570777 [Accessed November 30, 2016].

Grabski K, Lamalle L, Vilain C, Schwartz J-L, Vallée N, Tropres I, Baciu M, Le Bas J-F, Sato M (2012) Functional MRI assessment of orofacial articulators: neural correlates of lip, jaw, larynx, and tongue movements. Hum Brain Mapp 33:2306–2321 Available at: http://www.ncbi.nlm.nih.gov/pubmed/21826760 [Accessed October 21, 2013].

Grau C, Ginhoux R, Riera A, Nguyen TL, Chauvat H, Berg M, Amengual JL, Pascual-Leone A, Ruffini G (2014) Conscious Brain-to-Brain Communication in Humans Using Non-Invasive Technologies. PLoS One 9:e105225 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25137064 [Accessed August 20, 2014].

Grimaldi M, Fivela BG (2008) New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph. Proc … Available at: http://pasa.lira.dist.unige.it/pasapdf/332_Grimaldi_etal2008.pdf.

Grinvald A, Bonhoeffer T (1999) Optical imaging of electrical activity based on intrinsic signals and on voltage sensitive dyes: The methodology. Brain:1285–1366.

Grübler G (2011) Beyond the responsibility gap. Discussion note on responsibility and liability in the use of brain-computer interfaces. Ai Soc 26:377–382 Available at: http://link.springer.com/10.1007/s00146-011-0321-y [Accessed February 6, 2015].

Guenther FH (2006) Cortical interactions underlying the production of speech sounds. J Commun Disord 39:350–365 Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16887139.

Guenther FH, Brumberg JS, Wright EJ, Nieto-Castanon A, Tourville J a, Panko M, Law R, Siebert S a, Bartels JL, Andreasen DS, Ehirim P, Mao H, Kennedy PR (2009) A wireless brain-machine interface for real-time speech synthesis. PLoS One 4:e8218 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2784218&tool=pmcentrez&rendertype=abstract [Accessed September 27, 2013].

Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa O V. (1993) Magnetoencephalography - theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 65.

Hardcastle WJ, Roach PJ (1979) An instrumental investigation of coarticulation in stop consonant sequences, John Benja. Amsterdam.

Hasegawa T, Ohtani K (1992) Oral image to voice converter-image input microphone. [Proceedings] Singapore ICCS/ISITA `92:617–620.

Haselager P, Vlek R, Hill J, Nijboer F (2009) A note on ethical aspects of BCI. Neural Netw 22:1352–1357 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19616405 [Accessed January 19, 2015].

Heracleous P, Heracleous P, Nakajima Y, Nakajima Y, Lee A, Lee A, Saruwatari H, Saruwatari H, Shikano K, Shikano K (2004) Non-audible murmur (nam) speech recognition using a stethoscopic nam microphone. Proc ICLP:1469–1472.

Herff C, Heger D, de Pesters A, Telaar D, Brunner P, Schalk G, Schultz T (2015) Brain-to-text: decoding spoken phrases from phone representations in the brain. Front Neurosci 9:1–11 Available at: http://journal.frontiersin.org/article/10.3389/fnins.2015.00217 [Accessed June 16, 2015].

Herff C, Heger D, Putze F, Guan C, Schultz T (2012) Cross-subject classification of speaking modes using fNIRS. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 7664 LNCS:417–424.

Herff C, Johnson G, Diener L, Shih J, Krusienski D (n.d.) Towards direct speech synthesis from ECoG: A pilot study. CslUni-BremenDe:0–3 Available at: http://www.csl.uni-bremen.de/cms/images/documents/publications/HerffEMBC_16.pdf.

Hickok G, Houde J, Rong F (2011) Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69:407–422 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3057382&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2013].

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

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Hinterberger T, Kübler A, Kaiser J, Neumann N, Birbaumer N (2003) A brain-computer interface (BCI) for the locked-in: comparison of different EEG classifications for the thought translation device. Clin Neurophysiol 114:416–425 Available at: http://www.ncbi.nlm.nih.gov/pubmed/12705422 [Accessed December 7, 2016].

Hinton G (2010a) A Practical Guide to Training Restricted Boltzmann Machines. Comput Sci 7700:599–619.

Hinton G (2014) Dropout : A Simple Way to Prevent Neural Networks from Overfitting. 15:1929–1958.

Hinton GE (2010b) Rectified Linear Units Improve Restricted Boltzmann Machines. In: 27th International Conference on Machine Learning. Haifa, Israel.

Hinton GE, Osindero S (2006) A fast learning algorithm for deep belief nets ∗ 500 units 500 units.

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16873662 [Accessed October 17, 2013].

Hiroya S, Honda M (2004) Estimation of Articulatory Movements From Speech Acoustics Using an HMM-Based Speech Production Model. IEEE Trans Speech Audio Process 12:175–185 Available at: http://ieeexplore.ieee.org/document/1284345/ [Accessed February 25, 2017].

Hochberg L, Bacher D, Jarosiewicz B, Masse N, Simeral J, Vogel J, Haddadin S, Liu J, Cash S, van der Smagt P, Donoghue J (2012) Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485:372–375 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3640850&tool=pmcentrez&rendertype=abstract [Accessed September 16, 2013].

Hochberg L, Serruya M, Friehs G, Mukand J, Saleh M, Caplan A, Branner A, Chen D, Penn R, Donoghue J (2006) Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442:164–171 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16838014 [Accessed September 18, 2013].

Horikawa T, Tamaki M, Miyawaki Y, Kamitani Y (2013) Neural Decoding of Visual Imagery During Sleep. Science (80-) 340.

Hothorn T, Bretz F, Westfall P (2008) Simultaneous Inference in General Parametric Models. Biometrical J 50:346–363.

Huang J, Carr TH, Cao Y (2001) Comparing Cortical Activations for Silent and Overt Speech Using Event-Related fMRI. Hum Brain Mapp 15:39–53.

Huang T, Peng H, Zhang K (2013) Model Selection for Gaussian Mixture Models. :1–27 Available at: http://arxiv.org/abs/1301.3558.

Hueber T (2009) Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal : vers une communication parlée silencieuse.

Hueber T, Badin P, Savariaux C (2010a) Differences in articulatory strategies between silent, whispered and normal speech? a pilot study using electromagnetic articulography. Proc ISSP, to …:0–1 Available at: http://www.gipsa-lab.inpg.fr/~thomas.hueber/mes_documents/hueber_etal_issp2011.pdf.

Hueber T, Bailly G (2016) Statistical conversion of silent articulation into audible speech using full-covariance HMM. Comput Speech Lang 36:274–293 Available at:

http://www.sciencedirect.com/science/article/pii/S0885230815000340 [Accessed December 19, 2015].

Hueber T, Bailly G, Denby B (2012) Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface. Proc Interspeech:723–726 Available at: http://hal.archives-ouvertes.fr/hal-00741682/.

Hueber T, Benaroya E, Chollet G, Denby B, Dreyfus G, Stone M, Paristech PE, France PC, Pierre U, Paris C, France PC, Traitement L, Information D, Paristech T, France PC (2009) Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface Laboratoire d ' Electronique , Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Vocal Tract Visualization Lab , U. :640–643.

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

Hunt AJ, Black AW (1996) Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. Proc ICASSP:373–376.

Huo X, Wang J, Ghovanloo M (2008) A magneto-inductive sensor based wireless tongue-computer interface. IEEE Trans Neural Syst Rehabil Eng 16:497–504.

Ifft PJ, Shokur S, Li Z, Lebedev M a, International LS (2013) Brain-Machine Interface Enables Bimanual Arm Movements in Monkeys. Sci Transl Med 5.

Imai S (1983) Cepstral analysis synthesis on the mel frequency scale. Acoust Speech, Signal Process IEEE Int Conf ICASSP:93–96.

Imai S, Sumita K, Furuichi C (1983) Mel Log Spectrum Approximation (MLSA) Filter for Speech Synthesis. Electron Commun Japan 66-A:10–18.

Jackson A, Mavoori J, Fetz EE (2006) Long-term motor cortex plasticity induced by an electronic neural implant. Nature 444:56–60 Available at: http://www.nature.com/doifinder/10.1038/nature05226 [Accessed November 30, 2016].

Janke M, Wand M, Nakamura K, Schultz T (2012) Further investigations on EMG-to-speech conversion. ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc:365–368.

Janke M, Wand M, Schultz T (2010) Impact of lack of acoustic feedback in EMG-based silent speech recognition. Proc Interspeech:2686–2689 Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.187.2982&rep=rep1&type=pdf.

Jarosiewicz B, Bacher D, Sarma a a, Masse NY, Simeral JD, Sorice B, Oakley EM, Blabe CH, Pandarinath C, Cash SS, Eskandar E, Friehs G, Shenoy K V, Henderson JM, Donoghue JP, Hochberg LR (2015) Virtual typing by people with tetraplegia using a stabilized, self-calibrating intracortical brain-computer interface. IEEE BRAIN Gd Challenges Conf Washington, DC 7:1–11.

Jasmin KM, McGettigan C, Agnew ZK, Lavan N, Josephs O, Cummins F, Scott SK (2016) Cohesion and Joint Speech: Right Hemisphere Contributions to Synchronized Vocal Production. J Neurosci 36:4669–4680.

Jobsis F (1977) Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science (80-) 198:1264–1267.

Jorfi M, Skousen JL, Weder C, Capadona JR (2015) Progress towards biocompatible intracortical microelectrodes for neural interfacing applications. J Neural Eng 12:011001 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25460808 [Accessed December 7, 2016].

Jorgensen C, Dusan S (2010) Speech interfaces based upon surface electromyography. Speech Commun 52:354–366 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167639309001721 [Accessed June 25, 2015].

Jorgensen C, Lee DD, Agabon S (2003) Sub Auditory Speech Recognition Based on EMG/EPG Signals. Neural Networks 4:3128–3133.

Joseph S. Perkell (1974) A physiologically-oriented model of tongue activity in speech production. Available at: https://www.researchgate.net/publication/37991427_A_physiologically-oriented_model_of_tongue_activity_in_speech_production [Accessed February 15, 2017].

Jou S-C, Maier-Hein L, Schultz T, Waibel A (2006) Articulatory Feature Classification Using Surface Electromyography. Int Conf Acoust Speech, Signal Process:605–608.

Jou SCS, Schultz T, Waibel A (2007) Continuous electromyographic speech recognition with a multi-stream decoding architecture. ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc 4:401–404.

Kagawa Y, Shimoyama R, Yamabuchi T, Murai T, Takarada K (1992) Boundary element models of the vocal tract and radiation field and their response characteristics. 157:385–403 Available at: http://www.sciencedirect.com/science/article/pii/0022460X9290523Z.

Kamada K, Ogawa H, Kapeller C, Prueckl R, Guger C (2014) Rapid and low-invasive functional brain mapping by realtime visualization of high gamma activity for awake craniotomy. Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2014:6802–6805.

Kanas VG, Mporas I, Benz HL, Sgarbas KN, Bezerianos A, Nathan E (2014) Real-Time Voice Activity Detection for ECoG-Based Speech Brain Machine Interfaces. :862–865.

Kant E (1785) Fondements de la métaphysique des murs.

Käthner I, Wriessnegger SC, Müller-Putz GR, Kübler A, Halder S (2014) Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain-computer interface. Biol Psychol 102:118–129.

Kawahara H (1997) Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. 1997 IEEE Int Conf Acoust Speech, Signal Process 2:1303–1306.

Keller C, Kell CA (2016) Asymmetric intra- and interhemispheric interactions during covert and overt sentence reading. Neuropsychologia:1–18.

Kellis S, Miller K, Thomson K, Brown R, House P, Greger B (2010) Decoding spoken words using local field potentials recorded from the cortical surface. J Neural Eng 7:056007 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2970568&tool=pmcentrez&rendertype=abstract [Accessed September 28, 2013].

Kello CT, Plaut DC (2004) A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. J Acoust Soc Am 116:2354 Available at: http://link.aip.org/link/JASMAN/v116/i4/p2354/s1&Agg=doi [Accessed March 13, 2014].

Kelly JL, Lochbaum CC (1962) Speech synthesis. In: Proc. Fourth Int. Congress on Acoustics, pp 1–4. Copenhagen, Denmark.

Kenyon J (1929) The International Phonetic Alphabet. Am Speech:2005 Available at: http://www.jstor.org/stable/10.2307/452075.

Kipke DR, Shain W, Buzsáki G, Fetz E, Henderson JM, Hetke JF, Schalk G (2008) Advanced neurotechnologies for chronic neural interfaces: new horizons and clinical opportunities. J Neurosci 28:11830–11838 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3844837&tool=pmcentrez&rendertype=abstract [Accessed October 24, 2014].

Kiritani S (1986) X-ray microbeam method for measurement of articulatory dynamics-techniques and results. Speech Commun 5:119–140 Available at: http://linkinghub.elsevier.com/retrieve/pii/0167639386900038.

Kornhuber HH, Deecke L (2016) Brain potential changes in voluntary and passive movements in humans: readiness potential and reafferent potentials. Pflugers Arch Eur J Physiol 468:1115–1124 Available at: http://dx.doi.org/10.1007/s00424-016-1852-3.

Korning PG (1995) Training neural networks by means of genetic algorithms working on very long chromosomes. Int J Neural Syst 6:299–316 Available at: http://www.ncbi.nlm.nih.gov/pubmed/8589866.

Koskinen M, Viinikanoja J, Kurimo M, Klami A, Kaski S, Hari R (2013) Identifying fragments of natural speech from the listener's MEG signals. Hum Brain Mapp 34:1477–1489.

Kotchetkov IS, Hwang BY, Appelboom G, Kellner CP, Connolly ES (2010) Brain-computer interfaces: military, neurosurgical, and ethical perspective. Neurosurg Focus 28:E25 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20568942 [Accessed February 6, 2015].

Kourbeti I, Vakis A, Ziakas P, Karabetsos D, Potolidis E, Christou S, Samonis G (2015) Infections in patients undergoing craniotomy: risk factors associated with post-craniotomy meningitis. J Neurosurg 122:1113–1119.

Kubichek RF (1993) MEL-CEPSTRAL DISTANCE MEASURE FOR OBJECTIVE. :125–128.

Kuwabara H (1996) Acoustic properties of phonemes in continuous speech for different speaking rate. Proceeding Fourth Int Conf Spok Lang Process ICSLP '96 4.

Lachaux J, Jerbi K, Bertrand O, Minotti L, Hoffmann D, Schoendorff B, Kahane P (2007) BrainTV : a novel approach for online mapping of human. Biol Res:401–413.

Lachaux JP, Jung J, Mainy N, Dreher JC, Bertrand O, Baciu M, Minotti L, Hoffmann D, Kahane P (2008) Silence is golden: Transient neural deactivation in the prefrontal cortex during attentive reading. Cereb Cortex 18:443–450.

Laprie Y, Berger M-O (2015) Extraction of tongue contours in X-ray images with minimal user interaction. Proceeding Fourth Int Conf Spok Lang Process ICSLP '96 1:268–271 Available at:

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=607097 [Accessed June 25, 2015].

Laroche J, Stylianou Y, Moulines E (1993) HNS: Speech modification based on a harmonic+noise model. IEEE Int Conf Acoust Speech, Signal Process 2:550–553.

Lebedev MA, Tate AJ, Hanson TL, Li Z, O'Doherty JE, Winans JA, Ifft PJ, Zhuang KZ, Fitzsimmons NA, Schwarz DA, Fuller AM, An JH, Nicolelis MAL (2011) Future developments in brain-machine interface research. Clinics (Sao Paulo) 66 Suppl 1:25–32 Available at: http://www.ncbi.nlm.nih.gov/pubmed/21779720 [Accessed November 30, 2016].

Lecun Y, Denker JS, Solla SA (1990) Optimal Brain Damage. Adv Neural Inf Process Syst.

Lee D, Park B, Jang C, Park H-J (2011) Decoding brain states using functional magnetic resonance imaging. Biomed Eng Lett 1:82–88 Available at: http://link.springer.com/10.1007/s13534-011-0021-z [Accessed December 6, 2016].

Legnani FG, Saladino A, Casali C, Vetrano IG, Varisco M, Mattei L, Prada F, Perin A, Mangraviti A, Solero CL, DiMeco F (2013) Craniotomy vs. craniectomy for posterior fossa tumors: a prospective study to evaluate complications after surgery. Acta Neurochir (Wien) 155:2281–2286 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24078114 [Accessed November 30, 2016].

Lembke G, Erné SN, Nowak H, Menhorn B, Pasquarelli a (2014) Optical multichannel room temperature magnetic field imaging system for clinical application. Biomed Opt Express 5:876–881 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3959851&tool=pmcentrez&rendertype=abstract [Accessed December 3, 2014].

Lengellé R, Denoeux T (1996) Training MLPs Layer by Layer Using an Objective Function for Internal Representations. 9:83–97.

Leuthardt EC, Gaona C, Sharma M, Szrama N, Roland J, Freudenberg Z, Solis J, Breshears J, Schalk G (2011) Using the electrocorticographic speech network to control a brain-computer interface in humans. J Neural Eng 8:036004 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3701859&tool=pmcentrez&rendertype=abstract [Accessed September 26, 2013].

Lidz CW, Appelbaum PS, Joffe S, Albert K, Rosenbaum J, Simon L (n.d.) Competing commitments in clinical trials. IRB 31:1–6 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19873835 [Accessed November 30, 2016].

Liegeois-Chauvel C, de Graaf JB, Laguitton V, Chauvel P (1999) Specialization of left auditory cortex for speech perception in man depends on temporal coding. Cereb Cortex 9:484–496.

Lincoln M, Packman A, Onslow M (2006) Altered auditory feedback and the treatment of stuttering: A review. J Fluency Disord 31:71–89.

Ling ZH, Richmond K, Yamagishi J, Wang RH (2009) Integrating articulatory features into HMM-based parametric speech synthesis. IEEE Trans Audio, Speech Lang Process 17:1171–1185.

Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412:150–157 Available at: http://www.nature.com/doifinder/10.1038/35084005 [Accessed December 6, 2016].

Lotte F, Brumberg JS, Brunner P, Gunduz A, Ritaccio AL, Guan C, Schalk G (2015) Electrocorticographic representations of segmental features in continuous speech. Front Hum Neurosci 09:1–13 Available at: http://www.frontiersin.org/Human_Neuroscience/10.3389/fnhum.2015.00097/abstract.

Lowe DG (1999) Object recognition from local scale-invariant features. Proc Seventh IEEE Int Conf Comput Vis 2:1150–1157 Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=790410.

Lucivero F, Tamburrini G (2008) Ethical monitoring of brain-machine interfaces. AI Soc 22:449–460 Available at: http://link.springer.com/10.1007/s00146-007-0146-x [Accessed November 30, 2016].

Maaten L van der (n.d.) Matlab Toolbox for Dimensionality Reduction. http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html.

Maeda S (1982) A digital simulation method of the vocal-tract system. Speech Commun:199–229 Available at: http://www.sciencedirect.com/science/article/pii/0167639382900176.

Maeda S (1990) Compensatoy Articulation During Speech: Evidence from the analysis and Synthesis of Vocal-Tract shapes using an articulatory model. Speech Prod Speech Model:131–149.

Maier-Hein L (2005) Speech Recognition Using Surface Electromyography.

Maier-Hein L, Metze F, Schultz T, Waibel A (2005) Session independent non-audible speech recognition using surface electromyography. … Speech Recognit …:331–336 Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1566521.

Margalit E, Maia M, Weiland JD, Greenberg RJ, Fujii GY, Torres G, Piyathaisere D V, O'Hearn TM, Liu W, Lazzi G, Dagnelie G, Scribner DA, de Juan E, Humayun MS (2002) Retinal prosthesis for the blind. Surv Ophthalmol 47:335–356 Available at: http://www.ncbi.nlm.nih.gov/pubmed/12161210 [Accessed November 30, 2016].

Markel JD, Gray AH (1976) Linear Prediction of Speech. Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: http://link.springer.com/10.1007/978-3-642-66286-7 [Accessed November 13, 2015].

Martin S, Brunner P, Holdgraf C, Heinze H-J, Crone NE, Rieger J, Schalk G, Knight RT, Pasley BN (2014) Decoding spectrotemporal features of overt and covert speech from the human cortex. Front Neuroeng 7:14 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4034498&tool=pmcentrez&rendertype=abstract [Accessed July 24, 2014].

Mellinger J, Schalk G, Braun C, Preissl H, Rosenstiel W, Birbaumer N, Kübler A (2007) An MEG-based brain-computer interface (BCI). Neuroimage 36:581–593 Available at: http://www.ncbi.nlm.nih.gov/pubmed/17475511 [Accessed December 6, 2016].

Mercier-ganady J, Lotte F, Loup-escande E, Anatole L, Marchal M (2014) The Mind-Mirror : See Your Brain in Action in Your Head Using EEG and Augmented Reality. Virtual Real:33–38.

Mesgarani N, David S V, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am 123:899–909 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18247893 [Accessed November 30, 2016].

Middendorf M, McMillan G, Calhoun G, Jones KS (2000) Brain-computer interfaces based on the steady-state visual-evoked response. IEEE Trans Rehabil Eng 8:211–214.

Mill JS (1863) L'utilitarisme.

Montana DJ, Davis L (1989) Training Feedforward Neural Networks Using Genetic Algorithms. :762–767.

Morillon B, Liégeois-Chauvel C, Arnal LH, Bénar CG, Giraud AL (2012) Asymmetric function of theta and gamma activity in syllable processing: An intra-cortical study. Front Psychol 3:1–9.

Moritz CT, Perlmutter SI, Fetz EE (2008) Direct control of paralysed muscles by cortical neurons. Nature 456:639–642 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159518&tool=pmcentrez&rendertype=abstract [Accessed January 19, 2015].

Mugler EM, Patton JL, Flint RD, Wright Z a, Schuele SU, Rosenow J, Shih JJ, Krusienski DJ, Slutzky MW (2014) Direct classification of all American English phonemes using signals from functional speech motor cortex. J Neural Eng 11:035015 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24836588 [Accessed May 28, 2014].

Müller K-R, Tangermann M, Dornhege G, Krauledat M, Curio G, Blankertz B (2008) Machine learning for real-time single-trial EEG-analysis: from brain-computer interfacing to mental state monitoring. J Neurosci Methods 167:82–90 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18031824 [Accessed November 30, 2016].

Narasimha M, Peterson A (1978) On the Computation of the Discrete Cosine Transform. IEEE Trans Commun 26:934–936 Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1094144 [Accessed May 24, 2015].

Neerdael D (2015) Interfaces cerveau-machine. In: Encyclopédie du trans/posththumanisme, L'humain et ses préfixes, Vrin. (Hottois G, Missa J-N, Perdal L, eds), pp 388–397. Paris.

Ngoc YPT, Badin P (1994) Vocal tract acoustic transfer function measurements: further developments and applications. J Phys 4:549–552.

Nguyen D, Widrow B (1990) Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. IJCNN Int Jt Conf Neural Networks 13:C21.

Nijboer F, Clausen J, Allison BZ, Haselager P (2013) The Asilomar Survey : Stakeholders ' Opinions on Ethical Issues Related to Brain-Computer Interfacing. :541–578.

Nowakowska W, Gubrynowicz R, Zarnecki P (1993) On the model of vocal tract dynamics.

Nukada A (2006) FFTSS: A HIGH PERFORMANCE FAST FOURIER TRANSFORM LIBRARY. In: Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), pp 980–983.

Ogawa S, Lee TM, Kay a R, Tank DW (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci U S A 87:9868–9872 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=55275&tool=pmcentrez&rendertype=abstract.

Oord A Van Den, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) WaveNet: A Generative Model for Raw Audio. :1–15 Available at: http://arxiv.org/abs/1609.03499.

Osowski S (1993) New approach to selection of initial values of weights in neural function approximation. Electron Lett 29:313–315.

Palmer ED, Rosen HJ, Ojemann JG, Buckner RL, Kelley WM, Petersen SE (2001) An Event-Related fMRI Study of Overt and Covert Word Stem Completion. Neuroimage 14:182–193.

Pasley BN, David S V, Mesgarani N, Flinker A, Shamma S a, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10:e1001251 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3269422&tool=pmcentrez&rendertype=abstract [Accessed December 14, 2013].

Pasley BN, Knight RT (2013) Decoding Speech for Understanding and Treating Aphasia. Prog Brain Res 207:435–456.

Paul ES, Harding EJ, Mendl M (2005) Measuring emotional processes in animals: the utility of a cognitive approach. Neurosci Biobehav Rev 29:469–491.

Payan Y, Perrier P (1997) Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. Speech Commun 22:185–205.

Pei X, Barbour DL, Leuthardt EC, Schalk G (2011a) Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. J Neural Eng 8:046028 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3772685&tool=pmcentrez&rendertype=abstract [Accessed September 20, 2013].

Pei X, Leuthardt EC, Gaona CM, Brunner P, Wolpaw JR, Schalk G (2011b) Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. Neuroimage 54:2960–2972 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3020260&tool=pmcentrez&rendertype=abstract [Accessed May 27, 2014].

Penfield W, Boldrey E (1937) Somatic Motor and Sensory Representation in the Cerebral Cortex of Man as Studied by Electrical Stimulation. Brain 60:389–443.

Perrier P (2005) Control and representations in speech production. ZAS Pap Lingustics 40:109–132 Available at: http://www.zas.gwz-berlin.de/fileadmin/material/ZASPiL_Volltexte/zp40/zaspil40-perrier.pdf\nhttp://hal.archives-ouvertes.fr/hal-00430387.

Perrier P, Payan Y, Buchaillard S, Nazari MA, Chabanas M (2011) Biomechanical models to study speech. Faits de Langues 37:155–171.

Perrin F, Pernier J, Bertnard O, Giard MH, Echallier JF (1987) Mapping of scalp potentials by surface spline interpolation. Electroencephalogr Clin Neurophysiol 66:75–81.

Perrone-Bertolotti M, Rapin L, Lachaux JP, Baciu M, Lœvenbruck H (2014) What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. Behav Brain Res 261:220–239 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24412278.

Petersen SE, Fox PT, Posner MI, Mintun M, Raichle ME (1988) Positron emission tomographic studies of the cortical anatomy of single- word processing. Nature 331:585–589.

Petersen SE, Fox PT, Posner MI, Mintun M, Raichle ME (1989) Positron emission tomographic studies of the processing of single words. J Cogn Neurosci 1:153–170.

Pino A, Kalogeros E, Salemis E, Kouroupetroglou G (2003) Brain Computer Interface Cursor Measures for Motion- impaired and Able-bodied Users. Proc HCI Int 2003 10th Int Conf Human-Computer Interact Interact 4:1462–1466 Available at: http://speech.di.uoa.gr/sppages/spppdf/Final BCI HCII2003 _web_.pdf.

Polak E, Ribiere G (1969) Note sur la convergence de méthodes de directions conjuguées. ESAIM Math Model Numer Anal 3:35–43 Available at: https://eudml.org/doc/193115.

Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003) Recent advances in the automatic recognition of audio-visual speech. In: Proceedings of the IEEE, pp 1306–1326.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1988) Conjugate Gradient Methods in Multidimensions. In: Numerical recipes in C: The art of scientific computing, pp 420–425.

Price CJ, Wise RJ, Watson JD, Patterson K, Howard D, Frackowiak RS (1994) Brain activity during reading. The effects of exposure duration and task. Brain 117:1255–1269.

Pulvermüller F, Huss M, Kherif F, Moscoso del Prado Martin F, Hauk O, Shtyrov Y (2006) Motor cortex maps articulatory features of speech sounds. Proc Natl Acad Sci U S A 103:7865–7870 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1472536&tool=pmcentrez&rendertype=abstract.

Quade D (1979) Using Weighted Rankings in the Analysis of Complete Blocks with Additive Block Effects. J Am Stat Assoc 74:680–683 Available at: http://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481670 [Accessed March 19, 2014].

Quandt F, Reichert C, Hinrichs H, Heinze HJ, Knight RT, Rieger JW (2012) Single trial discrimination of individual finger movements on one hand: a combined MEG and EEG study. Neuroimage 59:3316–3324 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22155040 [Accessed December 6, 2016].

Rapin L, Dohen M, Polosan M, Perrier P, Lœvenbruck H (2013) An EMG study of the lip muscles during covert auditory verbal hallucinations in schizophrenia. J Speech Lang Hear Res 56:S1882–S1893 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24687444.

Rasmussen CE (1996) Function minimization using conjugate gradients. 0:1–7.

Ray S, Maunsell JHR (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol 9.

Reznikova Z (2007) Animal Intelligence, Cambridge . Cambridge.

Richmond K, Hoole P, King S, Forum I (n.d.) Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus. Proc Interspeech:1–4.

Rissman J, Greely HT, Wagner AD (2010) Detecting individual memories through the neural decoding of memory states and past experience. Proc Natl Acad Sci U S A 107:9849–9854 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20457911 [Accessed November 30, 2016].

Rubin P, Baer T (1981) An articulatory synthesizer for perceptual research. J Acoust Soc Am 70:321.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536.

Russell GO (1928) The vowel, its psychological mechanism, as shown by x-ray. Columbus OH Ohio State Univ Press.

Ryding E, Bradvik B, Ingvar D (1996) Silent Speech Activates Prefrontal Cortical Regions Asymmetrically , as Well as Speech-Related Areas in the Dominant Hemisphere. 52:435–451.

Sahin NT, Pinker S, Cash SS, Schomer D, Halgren E (2009) Sequential processing of lexical, grammatical, and phonological information within Broca's area. Science 326:445–449.

Sakoe H, Chiba S (1978) Dynamic Programming Algorithm Optimization for Spoken Word Recognition.

Schroeder MR (1993) A brief history of synthetic speech. Speech Commun 13:231–237.

Sellers EW, Ryan DB, Hauser CK (2014) Noninvasive brain-computer interface enables communication after brainstem stroke. Sci Transl Med 6:257re7.

Shewchuk JR (1994) An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. Science (80-) 49:64 Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.418&rep=rep1&type=pdf\nhttp://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.

Shuster LI, Lemieux SK (2005) An fMRI investigation of covertly and overtly produced mono- and multisyllabic words. Brain Lang 93:20–31 Available at: http://www.ncbi.nlm.nih.gov/pubmed/15766765.

Silbert LJ, Honey CJ, Simony E, Poeppel D, Hasson U (2014) Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc Natl Acad Sci 111:E4687–E4696.

Sitaram R, Zhang H, Guan C, Thulasidas M, Hoshi Y, Ishikawa A, Shimizu K, Birbaumer N (2007) Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface. Neuroimage 34:1416–1427 Available at: http://www.ncbi.nlm.nih.gov/pubmed/17196832 [Accessed December 6, 2016].

Sondhi M, Schroeter J (1987) A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans Acoust 35.

Song C, Xu R, Hong B, Member I (2014) Decoding of Chinese phoneme clusters using ECoG. :1278–1281.

Sörös P, Sokoloff LG, Bose A, Mcintosh AR, Graham SJ, Stuss DT (2006) Clustered functional MRI of overt speech production. Neuroimage 32:376–387.

Steinschneider M, Nourski K V, Fishman YI (2013) Representation of speech in human auditory cortex: is it special? Hear Res 305:57–73.

Stone M, Shawker TH, Talbot TL, Rich AH (1988) Cross-sectional tongue shape during the production of vowels. J Acoust Soc Am 83:1586–1596.

Stone M, Stock G, Bunin K, Kumar K, Epstein M, Kambhamettu C, Li M, Parthasarathy V, Prince J (2007) Comparison of speech production in upright and supine position. J Acoust Soc Am 122:532 Available at: http://scitation.aip.org/content/asa/journal/jasa/122/1/10.1121/1.2715659 [Accessed June 25, 2015].

Stylianou Y, Cappé O, Moulines E (1998) Continuous probabilistic transform for voice conversion. IEEE Trans Speech Audio Process 6:131–142.

Sugie N, Tsunoda K (1985) A speech prosthesis employing a speech synthesizer--vowel discrimination from perioral muscle activities and vowel production. IEEE Trans Biomed Eng 32:485–490.

Svec JG, Vampola T, Laukkanen a M (2011) Finite element modelling of vocal tract changes after voice therapy. 5:77–88.

Takahashi J, Suezawa S, Hasegawa Y, Sankai Y (2011) Tongue motion-based operation of support system for paralyzed patients. IEEE Int Conf Rehabil Robot 2011:5975359 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22275563 [Accessed November 30, 2016].

Takemoto H, Mokhtari P, Kitamura T (2010) Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method. J Acoust Soc Am 128:3724–3738.

Takemoto H, Mokhtari P, Kitamura T (2014) Comparison of vocal tract transfer functions calculated using one-dimensional and three-dimensional acoustic simulation methods National Institute of Information and Communications Technology , Japan Faculty of Intelligence and Informatics , Konan Univers. :408–412.

Talwar SK, Xu S, Hawley ES, Weiss SA, Moxon KA, Chapin JK (2002) Behavioural neuroscience: Rat navigation guided by remote control. Nature 417:37–38 Available at: http://www.nature.com/doifinder/10.1038/417037a [Accessed November 30, 2016].

Tamura Y, Ogawa H, Kapeller C, Prueckl R, Takeuchi F, Anei R, Ritaccio A, Guger C, Kamada K (2016) Passive language mapping combining real-time oscillation analysis with cortico-cortical evoked potentials for awake craniotomy. J Neurosurg:1–9.

Tankus A, Fried I, Shoham S (2012) Structured neuronal encoding and decoding of human speech features. Nat Commun 3:1015 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22910361 [Accessed June 3, 2014].

Tate MC, Herbet G, Moritz-Gasser S, Tate JE, Duffau H (2014) Probabilistic map of critical functional regions of the human cerebral cortex: Broca's area revisited. Brain 137:2773–2782 Available at: http://www.brain.oxfordjournals.org/cgi/doi/10.1093/brain/awu168.

Tennison MN, Moreno JD (2012) Neuroscience, ethics, and national security: the state of the art. PLoS Biol 10:e1001289 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3308927&tool=pmcentrez&rendertype=abstract [Accessed February 6, 2015].

Thimm G, Luettin J (1999) Extraction of Articulators in X-Ray Image Sequences. In: Proceedings of Eurospeech. Hungary.

Thomson EE et al. (2013) Perceiving invisible light through a somatosensory cortical prosthesis. Nat Commun 4:1482 Available at: http://www.nature.com/doifinder/10.1038/ncomms2497 [Accessed November 30, 2016].

Toda T, Black AW, Tokuda K (2008) Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Commun 50:215–227 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167639307001495 [Accessed November 13, 2013].

Tokuda K, Oura K, Tamamori A, Sako S, Zen H, Nose T, Takahashi T, Yamagishi J, Nankaku Y (2014) Speech Signal Processing Toolkit (SPTK). http://sp-tk.sourceforge.net/.

Tokuda K, Yoshimura T, Masuko T (1998) Speech parameter generation algorithms for hmm-based speech synthesis. :1315–1318.

Toth AR, Wand M, Schultz T (2009) Synthesizing Speech from Electromyography using Voice Transformation Techniques. :652–655.

Townsend G, Platsko V (2016) Pushing the P300-based brain-computer interface beyond 100 bpm: extending performance guided constraints into the temporal domain. J Neural Eng 13:026024.

Toyoda G, Brown EC, Matsuzaki N, Kojima K, Nishida M, Asano E (2014) Electrocorticographic correlates of overt articulation of 44 English phonemes: intracranial recording in children with focal epilepsy. Clin Neurophysiol 125:1129–1137.

Treder MS, Blankertz B (2010) (C)overt attention and visual speller design in an ERP-based brain-computer interface. Behav Brain Funct 6:28 Available at: http://www.ncbi.nlm.nih.gov/pubmed/20509913 [Accessed November 30, 2016].

Tucker DM (1993) Spatial sampling of head electrical fields: the geodesic sensor net. Electroencephalogr Clin Neurophysiol 87:154–163 Available at: http://linkinghub.elsevier.com/retrieve/pii/001346949390121B.

Uecker M, Zhang S, Voit D, Karaus A, Merboldt KD, Frahm J (2010) Real-time MRI at a resolution of 20 ms. NMR Biomed 23:986–994.

Ulman YI, Cakar T, Yildiz G (2015) Ethical Issues in Neuromarketing: "I Consume, Therefore I am!". Sci Eng Ethics 21:1271–1284 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25150848 [Accessed November 30, 2016].

Uría B (2011) A Deep Belief Network for the Acoustic-Articulatory Inversion Mapping Problem. Available at:

http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Deep+Belief+Network+for+the+Acoustic-Articulatory+Inversion+Mapping+Problem+Benigno+Ur&#65533;a#0.

Uria B, Renals S, Richmond K (2011) A Deep Neural Network for Acoustic-Articulatory Speech Inversion. :1–9.

Uutela K, Hämäläinen M, Somersalo E (1999) Visualization of magnetoencephalographic data using minimum current estimates. Neuroimage 10:173–180 Available at: http://www.sciencedirect.com/science/article/pii/S1053811999904548.

Vlek RJ, Steines D, Szibbo D, Kübler A, Schneider M-J, Haselager P, Nijboer F (2012) Ethical issues in brain-computer interface research, development, and dissemination. J Neurol Phys Ther 36:94–99 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22592066 [Accessed November 11, 2014].

Wagner M (1981) Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algortihms. In: Proceedings of ICASSP ’81, pp 1156–1159.

Waldert S, Preissl H, Demandt E, Braun C, Birbaumer N, Aertsen A, Mehring C (2008) Hand Movement Direction Decoded from MEG and EEG. J Neurosci 28.

Walliczek M, Kraft F, Jou S-C (Stan), Schultz T, Waibel A (2006) Sub-Word Unit based Non-audible Speech Recognition using Surface Electromyography. Proc 9th ISCA Int Conf Spok Lang Process:1487–1490 Available at: http://new.csl.url.for.files/WalliczekSchultz_Interspeech2006.pdf.

Wand M, Koutník J, Schmidhuber J (2016) Lipreading with Long Short-Term Memory. PhD Propos 1:248–256 Available at: http://arxiv.org/abs/1011.1669\nhttp://dx.doi.org/10.1088/1751-8113/44/8/085201\nhttp://arxiv.org/abs/1601.08188.

Wand M, Schulte C, Janke M, Schultz T (2013) Array-based Electromyographic Silent Speech Interface. In: 6th International Conference on Bio-inspired Systems and Signal Processing.

Wand M, Schultz T (2011) Session-independent EMG-based Speech Recognition. Biosignals:295–300.

Wand M, Schultz T (2014) Towards real-life application of EMG-based speech recognition by using unsupervised adaptation. Proc Annu Conf Int Speech Commun Assoc INTERSPEECH:1189–1193.

Weiskopf N, Mathiak K, Bock SW, Scharnowski F, Veit R, Grodd W, Goebel R, Birbaumer N (2004) Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI). IEEE Trans Biomed Eng 51:966–970.

Wodlinger B, Downey JE, Tyler-Kabara EC, Schwartz a B, Boninger ML, Collinger JL (2014) Ten-dimensional anthropomorphic arm control in a human brain-machine interface: difficulties, solutions, and limitations. J Neural Eng 12:016011 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25514320 [Accessed December 17, 2014].

Wolfe P (1969) Convergence Conditions for Ascent Methods. SIAM Rev 11:226–235 Available at: http://epubs.siam.org/doi/abs/10.1137/1011036 [Accessed September 1, 2016].

Wolfe P (1971) Convergence Conditions for Ascent Methods. II: Some Corrections. SIAM Rev 13:185–188 Available at: http://epubs.siam.org/doi/abs/10.1137/1013035 [Accessed September 1, 2016].

Wolpe PR, Foster KR, Langleben DD (2005) Emerging neurotechnologies for lie-detection: promises and perils. Am J Bioeth 5:39–49 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16036700 [Accessed November 30, 2016].

Wu HC, Nagasawa T, Brown EC, Juhasz C, Rothermel R, Hoechstetter K, Shah A, Mittal S, Fuerst D, Sood S, Asano E (2011) Gamma-oscillations modulated by picture naming and word reading: Intracranial recording in epileptic patients. Clin Neurophysiol 122:1929–1942.

Yam JYF, Chow TWS (2000) A weight initialization method for improving training speed in feedforward neural network. Neurocomputing 30:219–232 Available at: http://linkinghub.elsevier.com/retrieve/pii/S0925231299001277.

Yoshimura N, Nishimoto A, Belkacem AN, Shin D, Kambara H, Hanakawa T, Koike Y (2016) Decoding of covert vowel articulation using electroencephalography cortical currents. Front Neurosci 10:1–15.

Young SJ, Evermann G, Gales MJF, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland PC (2009) The HTK Book (for HTK Version 3.4). Construction:384 Available at: http://htk.eng.cam.ac.uk.

Youssef A Ben (2011) Control of talking heads by acoustic-to-articulatory inversion for language learning and rehabilitation.

Ze H, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural networks. 2013 IEEE Int Conf Acoust Speech Signal Process:7962–7966 Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6639215.

Zeiler M, Fergus R (2012) Regularization of Neural Networks using DropConnect.

Zen H, Nankaku Y, Tokuda K (2011) Continuous Stochastic Feature Mapping Based on Trajectory HMMs. IEEE Trans Audio Speech Lang Processing 19:417–430 Available at: http://ieeexplore.ieee.org/document/5458026/ [Accessed February 25, 2017].

Zepeda A, Arias C, Sengpiel F (2004) Optical imaging of intrinsic signals: Recent developments in the methodology and its applications. J Neurosci Methods 136:1–21.

Zhu Y, Toutios A, Narayanan S, Nayak K (2013) Faster 3D Vocal Tract Real - time MRI Using Constrained Reconstruction. :1292–1296.

Ziolko B, Ziolko M (2011) Time Durations of Phonemes in the Polish Language. Comput Sci 6562:105–114.

Publications

Journal articles

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

Bocquelet F., Hueber T., Girin L., Chabardès S., Yvert B. (2016) **Key considerations in designing a speech brain-computer interface.** Journal of Physiology, in revision.

Book chapters

Bocquelet F., Piret G., Aumonier N., Yvert B. (2015) **Ethique et interfaces cerveau-ordinateur.** In Bongrain L., Clerc M., Lotte F. **Interfaces cerveau-ordinateur : méthodes, applications et perspectives,** ISTE-Wiley, 2015.

International conferences

Bocquelet F., Abdoun O., Joucla S., Yvert B. (2013) **Explore your MEA data with NeuroMap.** Fourth GDR 2904 Conference on Neuronal Ensemble Recordings in Integrative Neuroscience. Bordeaux, Oct 10, France.

Bocquelet F., Hueber T., Girin L., Badin P., Yvert B. (2014) **Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications.** Proceedings of the Interspeech Conference, Singapore, Sept 14-18.

Bocquelet F., Hueber T., Savariaux C., Girin L., Yvert B. (2015) **Real-time control of an articulatory speech synthesizer for BCI applications.** IEEE Eng Med Biol Soc Conference, Montpellier, France, April 22-24.

Bocquelet F., Hueber T., Girin L., Badin P., Yvert B. (2015) **Real-time Control of a DNN-based Articulatory Synthesizer for Silent Speech Conversion: a pilot study.** Proceedings of the Interspeech Conference, Dresden, Germany, Sept 6-10.

Fabre D., Hueber T., Bocquelet F., Badin P. (2015) **Tongue Tracking in Ultrasound Images using EigenTongue Decomposition and Artificial Neural Networks.** Proceedings of the Interspeech Conference, Dresden, Germany, Sept 6-10.

Bocquelet F., Hueber T., Girin L., Savariaux C., Yvert B. (2015) **Real-time articulatory speech synthesis for brain-computer interfaces.** SFN 2015, Chicago, Illinois, USA, Oct 17-21.

Résumé en français

I. Introduction

En France, environ 300 000 personnes souffrent d'un trouble important de la parole ou d'une aphasie souvent due à un accident cardio-vasculaire, mais aussi à certaines paralysies sévères, au syndrome d'enfermement, à des maladies neurodégénératives comme la sclérose latérale amyotrophique ou la maladie de Parkinson, à des myopathies ou encore à un coma. Certains patients sont dans l'impossibilité totale de communiquer alors que leurs capacités cognitives et sensorielles sont préservées. Pour ces personnes, la perte de la parole est une affliction supplémentaire qui empire leur condition : elle rend la communication avec le personnel soignant difficile, et peut entrainer l'isolement social ou la dépression. Par conséquent, il est crucial pour ces patients de récupérer leur capacité à communiquer avec le monde extérieur.

Certaines approches actuelles peuvent fournir un moyen de communiquer, principalement via des dispositifs de saisie lettre par lettre qui exploitent des mouvements résiduels des yeux, ou les réponses cérébrales à certains stimuli spécifiques. Cependant, plusieurs minutes sont requises pour taper une phrase complète, alors que seulement quelques secondes sont nécessaires en utilisant la parole naturelle ; et tous les patients ne peuvent bénéficier de ce type de système.

La parole reste en effet notre moyen de communication le plus naturel et efficace. Mais elle est aussi le résultat de mouvements musculaires complexes, contrôlés par notre système nerveux. Au cours des dernières décennies, des approches utilisant des interfaces cerveau-machine (BCI, pour « Brain-computer interface ») ont successivement été développées afin de contrôler le mouvement d'effecteurs (par exemple un bras robotisé ou une souris d'ordinateur), avec une précision grandissante, d'abord chez l'animal et plus récemment chez l'Homme. Ces systèmes ont tout d'abord permis le contrôle d'effecteurs avec peu de degrés de liberté, typiquement 1 ou 2, mais les études les plus récentes ont montré que des sujets pouvaient contrôler simultanément jusqu'à dix degrés de liberté après un entrainement approprié, afin d'achever des taches motrices complexes. Jusqu'à ce jour il n'y a toutefois pas encore eu de démonstration quant à la faisabilité de restaurer la parole naturelle via une approche BCI.

L'objectif général de cette thèse était donc de franchir les premières étapes vers une telle preuve de concept. En particulier, son but principal était de développer un synthétiseur de parole paramétrique pouvant être utilisé dans un paradigme BCI (**Partie 3** de cette thèse), et de mettre en place des essais cliniques afin de collecter, analyser et décoder l'activité neuronale de la parole (**Partie 4**).

II. Résumé de l'état de l'art

1. Interfaces cerveau-machine pour la restauration de la parole

Plusieurs approches ont été proposées pour restaurer la communication chez les patients atteints de paralysies sévères, généralement via un processus de saisie lettre par lettre qui exploite des signaux physiologiques résiduels, par exemple en suivant la direction du regard pour contrôler la position d'un curseur sur un écran d'ordinateur, et en détectant les clignement d'œil pour actionner un clic. Ces solutions ne sont en revanches disponibles que pour les patients gardant un contrôle moteur suffisant, et permettent de ne contrôler que des effecteurs avec peu de degrés de liberté. Pour surmonter ces difficultés, des systèmes de communication directement contrôlés par les signaux neuronaux ont été proposés.

De tels systèmes exploitent le potentiel évoqué P300 enregistré par électro-encéphalographie (EEG). Le P300 est une réponse neuronale qui apparait généralement lorsqu'un évènement peu probable mais attendu survient pendant une série d'évènement très probables, par exemple lorsqu'un sujet détecte activement un son différent parmi une suite de sons identiques. De manière similaire, lorsque la rétine est exposée à une stimulation visuelle périodique, le cerveau génère un potentiel visuel évoqué oscillant à la même fréquence, aussi appelé potentiel « steady-state ». Lorsqu'un sujet porte successivement son attention sur les lettres de l'alphabet clignotant à des fréquences différentes sur un écran, ce potentiel peut être utilisé pour repérer chaque lettre et épeler ainsi une phrase complète. Ce type de méthode est cependant limité par une faible vitesse d'épellation de quelques caractères par minutes. De plus, ce type de tâche est particulièrement exhaustif pour le patient qui doit rester concentré pendant tout le processus d'épellation de la phrase, limitant ainsi son usage sur de longues périodes.

D'un autre côté, les systèmes BCIs basés sur des enregistrements intra-corticaux, bien qu'ayant l'inconvénient majeur d'être invasifs, semble requérir un effort de concentration moindre de la part du sujet, pour qui l'usage du système devient peu à peu naturel. Les enregistrements corticaux permettent par ailleurs de capturer plus d'information et conduisent donc à un meilleur décodage des intentions de leur utilisateur, ce qui conduit à une vitesse d'épellation bien supérieur, d'environ 20 à 30 caractères par minute, et pendant des durées plus longues.

En revanche, ce type de système exploite en général l'activité neuronale venant des aires corticales motrices de la main ou du bras et forment donc une manière indirecte de communiquer en exploitant des aires qui ne relèvent pas de la parole. En vue de décoder de la parole complète à partir de l'activité neuronale, il semble plus efficace d'enregistrer l'activité neuronale dans les zones corticales spécifiques à la production de la parole.

a. Les aires corticales de la production de la parole

Les aires corticales de la production de la parole, y compris de la parole imaginée, ont été largement étudiées en utilisant différentes techniques d'imagerie, allant de l'imagerie par résonance magnétique fonctionnelle (IRMf) à l'utilisation de microélectrodes intra-corticales. Dans l'ensemble, les résultats de ces différentes études suggèrent que la région gauche

inférieure frontale englobant les aires de Brodmann 4, 6, 43, 44 et 45 sont des candidats pertinents pour l'enregistrement et le décodage d'activité neuronale dans le but de contrôler un synthétiseur de parole. Cependant, ces études ont également montré qu'il existe une forte variabilité dans la localisation exacte de ces zones, probablement dues aux spécificités individuelles de chaque sujet, mais aussi aux différents types de tâches effectuées (parole libre, répétition d'un mot, imagination d'une voyelle, etc.) et aux méthodes utilisées pour l'enregistrement de l'activité neuronale. Par conséquent, bien qu'il y ait des preuves d'activation corticale dans ces aires durant la production de parole orale ou imaginée, ces résultats suggèrent également que ces zones spécifiques devraient être identifiées individuellement.

b. Décodage de la parole à partir de l'activité neuronale

Plusieurs études ont visé à décoder la parole ou ses caractéristiques à partir de l'activité neuronale. Ces approches de décodage peuvent être divisées en deux grandes catégories : les approches « discrètes », et les approches « continues ». Les approches discrètes visent à classifier l'activité neuronale en plusieurs catégories, généralement correspondant à des unités phonétiques comme des phones ou des mots ; alors que les approches continues ne se basent pas sur le décodage d'une représentation intermédiaire discrète mais prédisent plutôt des paramètres continus, comme les trajectoires acoustiques.

Les méthodes discrètes passant par une représentation intermédiaire phonétique, elles ont le principal avantage de pouvoir intégrer des connaissances linguistiques (par exemple un dictionnaire de mots prédéfinis ou des règles de grammaire) permettant d'améliorer fortement le décodage. En revanche, ces méthodes induisent généralement un délai additionnel entre l'intention de parole du patient et la synthèse effective de la parole prédite. Il est connu qu'un délai trop important (supérieur à 50ms) entre l'intention et le retour auditif perturbe généralement la production de la parole. D'un autre côté, les approches continues permettent une synthèse directe et sans délai, offrant donc un retour presque immédiat au sujet. Par ailleurs, plusieurs études sur les BCIs ont montré l'importance de l'entrainement du sujet pour améliorer la précision de son contrôle du système. Un retour immédiat semble plus propice à la réussite d'un tel entrainement. Pour ces raisons, nous avons choisi dans cette thèse de considérer un décodage continu de l'activité neuronale en parole.

De plus, alors que les études existantes ont principalement considéré le décodage de caractéristiques acoustiques de la parole, plusieurs études récentes ont montré que l'activité des aires frontales du cortex moteur de la parole reflétait plutôt ses propriétés articulatoires (par exemple l'ouverture des lèvres, l'avancement de la langue, etc.). Alors que l'activité neuronale pourrait être directement décodée en paramètres acoustiques, ces études soutiennent l'hypothèse qu'une stratégie pertinente pourrait être de considérer une approche plus « indirecte » dans laquelle les signaux corticaux sont décodés en caractéristiques articulatoires servant à contrôler un synthétiseur de parole articulatoire.

Dans cette thèse, nous avons considéré cette hypothèse. Afin de pouvoir synthétiser de la parole, l'un des prérequis est d'avoir à disposition un synthétiseur de parole qui convertit des trajectoires articulatoires (i.e. les mouvements des principaux articulateurs du conduit vocal) en un signal audible de parole.

Ces choix sont motivés plus en détail dans le **Chapitre 1** de cette thèse.

2. Synthèse de parole à partir de données articulatoires

La production de la parole peut être assimilée à la manière dont sont jouées les notes d'un instrument à vent. Le système respiratoire joue le rôle de la soufflerie en expulsant de l'air à travers la trachée. Le larynx, traversé par les cordes vocales, joue le rôle de l'appareil vibratoire en modulant périodiquement ce flux d'air pendant la production de sons voisés. Enfin, le conduit vocal joue le rôle de résonateur et transforme le flux d'air modulé en parole en modulant sa géométrie à l'aide d'organes mobiles, les articulateurs.

Il existe donc un lien direct entre la géométrie du conduit vocal et la parole produite. Il est donc possible de synthétiser de la parole à partir de trajectoires articulatoires, i.e. à partir des mouvements des articulateurs du conduit vocal. Dans ce domaine, deux types d'approches existent : les méthodes dites « physiques » qui visent à mimer de façon réaliste les propriétés acoustiques et géométriques du conduit vocal, via des simulations physiques de propagation d'onde, et les méthodes dites « statistiques » qui exploitent de larges bases de données articulatoires et acoustiques afin de modéliser la relation qui lie l'acoustique à l'articulatoire d'un point de vue probabiliste, sans chercher à en expliquer les mécanismes physiques. Dans cette thèse nous avons choisi de considérer une approche statistique (voir **Chapitre 2** pour plus de détails).

Tout comme pour le décodage de l'activité neuronale, les approches statistiques de synthèse de parole à partir de données articulatoires peuvent elles-mêmes être divisées en deux grandes catégories : les approches « discrètes », qui passent par une représentation intermédiaire phonétique, et les approches « continues », qui estiment directement l'acoustique à partir de l'articulatoire, sans passer par une représentation phonétique. Pour des raisons similaires à celles évoquées pour le décodage de la parole à partir de l'activité neuronale, nous avons choisi ici de considérer une approche continue de synthèse, approche qui avait par ailleurs déjà montré des résultats prometteurs.

Comme mentionné précédemment, les approches de synthèses statistiques exploitent de larges corpus de données acoustiques et articulatoires. Alors que l'acoustique est exclusivement enregistrée à l'aide de microphones, plusieurs méthodes différentes existent pour l'acquisition de données articulatoires : imagerie au rayon X, imagerie par résonance magnétique, imagerie à ultrasons, etc. Parmi l'ensemble de ces techniques, nous avons fait le choix d'utiliser l'électro-magnéto articulographie, qui permet d'enregistrer la position 3D de capteurs collés sur les principaux articulateurs avec une forte précision spatiale (inférieure au millimètre) et temporelle (de l'ordre de 400Hz), tout en étant peu risquée pour le patient (contrairement à l'utilisation de rayons X par exemple).

Ces considérations sont détaillées dans le **Chapitre 2** de cette thèse.

III. Synthèse de parole à partir de données articulatoires

1. *Enregistrement d'un corpus articulatoire*-acoustique

Les approches statistiques de synthèse de parole à partir de données articulatoires reposent sur l'utilisation de larges corpus contenant des données articulatoires et acoustiques synchrones. Bien qu'un tel corpus fût déjà existant, des résultats préliminaires ont suggéré que l'enregistrement d'un nouveau corpus plus complet pourrait fortement améliorer la qualité finale de la synthèse.

Une première étape de cette thèse a donc consisté à enregistrer un tel corpus, en utilisant la méthode de l'électro-magnéto articulographie (EMA) : de petites bobines sont collées sur les différents articulateurs du conduit vocal et placées dans un champ électromagnétique variable. En mesurant le courant induit dans les bobines, il est possible d'estimer leur position dans l'espace avec une grande précision spatiale et temporelle.

Le corpus articulato-acoustique final, appelé « BY2014 », a été enregistré chez un sujet masculin français. Une bobine a été placée sur chaque lèvre (supérieure et inférieure), une sur chaque commissure des lèvres (gauche et droite), trois sur la langue (avant, milieu, arrière), une sur la mâchoire, et une sur le palais mou. Une bobine supplémentaire, dont non seulement la position mais aussi l'orientation ont été enregistrées, a été utilisée afin de compenser les éventuels mouvements de tête du sujet. Le signal audio a lui été enregistré à l'aide d'un microphone, puis a été paramétré par 25 coefficients acoustiques, appelés coefficients mel-cepstraux. Cette paramétrisation permet d'obtenir une représentation du signal audio plus adaptée aux méthodes d'apprentissage automatique, et qui peut être facilement retransformée en un signal audible.

Au total, 925 segments de parole différents ont été enregistrés, incluant toutes les voyelles isolées, toutes les séquences voyelle-consone-voyelle démarrant et finissant par la même voyelle, ainsi que des phrases allant de phrases courtes (typiquement 4-5 mots) à de longues phrases issues de journaux d'actualité (typiquement 10-20 mots), ce qui représente environ 45 minutes de paroles une fois les périodes de silence retirées. Cette base de données a par ailleurs été diffusée publiquement.

Plus de détail sur le corpus et son enregistrement sont indiqués dans le **Chapitre 3** de cette thèse.

2. Synthèse de parole à partir de données articulatoires

Le corpus articulo-acoustique précédemment décrit a ensuite été utilisé pour calculer automatiquement (ou « entrainer ») un modèle mathématique permettant de transformer de nouvelles trajectoires articulatoires en paramètres acoustiques. Deux types de modèles mathématiques ont été comparés : les modèles de mixtures de gaussiennes (GMM pour « Gaussian Mixture Model »), et les réseaux de neurones profonds (DNN pour « Deep Neural Network »). Ici, les GMMs ont permis de modéliser la distribution de probabilité jointe des données acoustiques et articulatoires, puis cette distribution a été utilisée pour inférer les paramètres acoustiques correspondant à de nouvelles données articulatoires qui n'avaient pas

été précédemment observées. De façon similaire, les DNNs ont été entrainés afin d'estimer une fonction mathématique complexe permettant de passer des données articulatoires aux données acoustiques. Plus de détails sur les fondements théoriques de ces deux méthodes sont fournies à la fin du **Chapitre 2** de cette thèse, et une méthode spécifique à l'entrainement des DNNs a été proposée dans le **Chapitre 4**.

Comme nous l'avons précédemment motivé dans le **Chapitre 1** de cette thèse, un synthétiseur de parole conçu pour être utilisé par une interface cerveau-machine doit être contrôlable par aussi peu de paramètres que possible, et être relativement robuste aux fluctuations de ces paramètres. Afin d'évaluer ceci, les données articulatoires ont été artificiellement dégradées, en leur ajoutant du bruit et/ou en réduisant le nombre de paramètres articulatoires via différentes techniques de réduction de dimensionnalité.

La parole synthétisée a ensuite été évaluée de manière objective, en utilisant un système de reconnaissance automatique de la parole, et de manière subjective, en demandant à des sujets d'identifier les sons synthétisés. Les résultats ont tout d'abord montré que de la parole intelligible pouvait être synthétisée, en temps réel, avec un taux de reconnaissance sur les mots supérieur à 90% en utilisant 14 paramètres articulatoires. Ce nombre de paramètres a par ailleurs pu être réduit, tout en conservant une bonne intelligibilité avec environ 10 paramètres. Enfin, les résultats ont également suggéré que l'approche utilisant les DNNs serait plus robuste aux fluctuations des paramètres articulatoires que celle utilisant des GMMs.

Ces résultats sont détaillés dans le **Chapitre 4** de cette thèse.

3. Contrôle temps-réel du synthétiseur à partir de parole silencieuse

Bien que le synthétiseur de parole décrit dans la section précédente a pu produire de la parole intelligible, celui-ci a été conçu à partir et testé sur des données d'un locuteur unique, enregistrées au cours d'une unique session. Dans un second temps, nous avons donc cherché à savoir si un tel synthétiseur pouvait être contrôlé en temps-réel par un locuteur différent, ou par le même locuteur mais au cours d'une session différente. En effet, d'une session à l'autre, les capteurs EMA peuvent ne pas être placés exactement à la même position et avec la même orientation, ou encore le nombre de capteurs peut varier, ou le locuteur peut être un nouveau sujet avec une géométrie du conduit vocal différente et une manière différente d'articuler les sons. Un calibrage est donc nécessaire afin de compenser ces différences et de transformer l'espace articulatoire d'un nouveau locuteur en celui du locuteur de référence à partir duquel a été construit le synthétiseur.

Dans un premier temps, 50 phrases courtes issues du corpus articulo-acoustique de référence ont été présentées aux nouveaux locuteurs, trois fois de suite et à un rythme prédéfini. Chaque locuteur (4 au total) devait alors articuler de façon synchrone ces phrases, sans les prononcer à voix haute, alors que leurs trajectoires articulatoires étaient enregistrées par électro-magnéto articulographie. Ceci a permis d'obtenir pour ces 50 phrases les mouvements articulatoires des nouveaux sujets, synchronisés avec ceux du locuteur de référence. Un modèle linéaire a ensuite été entrainé sur ces données afin de transformer l'espace articulatoire des nouveaux locuteurs en celui du locuteur de référence.

Dans un second temps, ce modèle linéaire a été cascadé avec le synthétiseur de parole afin de permettre au sujet de contrôler, en temps réel, le synthétiseur à partir de parole silencieuse (i.e. en articulant mais sans prononcer à haute voix) alors qu'il recevait le retour de la synthèse via des écouteurs. Pendant cette période, chaque sujet devait articuler un ensemble prédéfini de voyelles et séquences voyelle-consone-voyelle. Ces éléments ont ensuite été évalués lors d'un test perceptif.

Les résultats ont montré que les nouveaux locuteurs pouvaient contrôler, en temps-réel, le synthétiseur, mais avec une intelligibilité moindre qu'en utilisant les données de référence. Par ailleurs, des épisodes de conversation spontanée ont pu avoir lieu pour deux des quatre sujets.

Ces résultats sont détaillés dans le **Chapitre 5** de cette thèse.

IV. Vers une interface cerveau-machine pour la restauration de la parole

1. Cartographie peropératoire des aires corticales de la parole

Au cours de ma thèse, nous avons collaboré avec le professeur Stéphan Chabardès, neurochirurgien au centre hospitalier universitaire de Grenoble. Ceci nous a permis de recueillir de l'activité neuronale chez des patients subissant une chirurgie éveillée pour le retrait d'une tumeur cérébrale. En effet, un des traitements en cas de tumeur cérébrale consiste à retirer les zones cancérigènes tout en préservant les fonctions sensori-motrices du patient. La localisation des aires fonctionnelles à préserver, comme celles de la parole, est donc cruciale. Cela peut se faire par exemple par imagerie par résonance magnétique fonctionnelle avant la chirurgie, mais avec une résolution spatiale limitée ; ou alors via des stimulations électriques délivrées à différentes positions du cortex pendant la chirurgie alors que le patient est éveillé et fournit un retour de sensations au chirurgien. Cependant, cette dernière méthode est indirecte, et la délivrance de stimulations électriques peut déclencher des crises d'épilepsie. Une autre solution pourrait être d'effectuer une cartographie fonctionnelle des aires corticales via des enregistrements électrophysiologiques effectués par des électrodes placées en surface du cortex.

Au cours de cette thèse, nous avons donc enregistré l'activité cérébrale de deux patients pour lesquels la tumeur était située proche du cortex moteur de la parole. Les enregistrements ont été effectués par électro-corticographie pendant que les patient parlaient à voix haute ou imaginaient parler, via une matrice d'électrodes placées en surface du cortex. Nous avons également développé une approche permettant de cartographier l'activité cérébrale spécifique à la parole, directement pendant la chirurgie éveillée, sur le champ opératoire du chirurgien. Cette cartographie pourrait être utilisée pour identifier les aires de la parole afin de les préserver pendant la résection de la tumeur. C'était également une bonne opportunité d'enregistrer des données neuronales pendant la production de la parole, afin d'étudier le décodage de la parole à partir de données neuronales.

Afin d'effectuer la cartographie de l'activité neuronale spécifique à la parole, un microphone était placé devant chaque patient. Les périodes de paroles et celles de silence ont été automatiquement détectées par une méthode simple fonctionnant en temps réel (voir

Chapitre 5 pour plus de détails). En parallèle, les signaux neuronaux ont été analysés afin d'en extraire des caractéristiques pertinentes. Pour cela, la représentation temps-fréquence de chaque signal a été calculée, et des tests statistiques ont permis d'automatiquement extraire, pour chaque électrode, les fréquences du signal présentant une différence significative entre les instants de parole et ceux de silence. Ces différences ont pu ensuite être visualisées sur des cartes d'activité alignées sur des photographies du champ opératoire du neurochirurgien, via des méthodes d'alignement et d'interpolation semi-automatiques.

Les résultats ont montré qu'une désynchronisation béta (diminution de l'activité neuronale pour les fréquences du signal entre 10 et 30 Hz) se produisait dans le cortex moteur pendant la production de parole, mais également pendant l'imagination de parole. Cette désynchronisation béta était également accompagnée d'une augmentation d'activité dans le gamma et haut-gamme (fréquences supérieures à 70Hz), également dans le cortex moteur, aussi bien pour la parole orale qu'imaginée. Ceci suggère que la parole orale et la parole imaginée partagent, au moins partiellement, une représentation neuronale commune dans le cortex moteur.

Ces résultats sont détaillés dans le **Chapitre 6** de cette thèse.

2. *Décodage de la parole à partir de l'activité corticale*

Dans un second temps, nous avons étudié le décodage de la parole à partir de l'activité neuronale des deux patients précédents. En particulier, nous avons considéré trois étapes clés du décodage de la parole : le décodage de l'intention de parler (volonté ou non de parler), le décodage du voisement (vibration ou non des cordes vocales pendant la parole) et le décodage des paramètres articulatoires ou acoustiques de la parole.

Ce décodage a été effectué en ne conservant, pour chaque patient, que les signaux provenant des électrodes situées sur des aires spécifiques à la parole obtenues par la cartographie précédente. Le décodage de l'intention de parler et du voisement ont été effectués à l'aide de machines à vecteurs de support (SVM pour « Support Vector Machine »), une technique d'apprentissage supervisé permettant d'effectuer de la classification de données. Le décodage des paramètres articulatoires a été effectué via des modèles linéaires, et comparé au décodage de paramètres acoustiques afin de tester l'hypothèse initiale de cette thèse.

Les résultats ont montré que l'intention de parler pouvait être décodée très largement au-dessus du niveau de la chance pour la parole orale, avec un taux de classification supérieur à 90% pour le second patient. L'intention de parler a également pu être décodée au-dessus de la chance pour le cas de la parole imaginée dans le cas du premier patient (le second patient n'ayant pas produit de parole imaginée). Le voisement lors de la production de parole a lui aussi pu être décodé au-dessus de la chance, mais dans une moindre mesure (de l'ordre de 75%). Enfin, les paramètres articulatoires et acoustiques ont eux aussi pu être décodés au-dessus de la chance, mais dans une moindre mesure, sans pouvoir produire de parole intelligible. Les résultats préliminaires ont par ailleurs montré un meilleur décodage des paramètres articulatoires comparé à celui des paramètres acoustiques, ce qui soutient l'hypothèse initiale de cette thèse. Ces résultats doivent cependant être confirmés par de futures expériences.

L'ensemble de ces résultats est décrit plus en détails dans le **Chapitre 7** de cette thèse.

V. Questions éthiques relatives aux interfaces cerveau-machine

Enfin, tout au long de cette thèse, nous nous sommes intéressés aux implications éthiques du développement des interfaces cerveau-machine. Cette réflexion a été conduite sur trois niveaux, concernant respectivement l'animal, l'Homme et l'humanité.

En particulier, nous nous sommes intéressés à la souffrance animale, à la gestion de l'espoir suscité chez les patients participant aux essais cliniques, à leur consentement éclairé et à la balance bénéfices/risques de ces essais. Mais également aux futurs cas d'usages des BCIs et leurs implications en terme de sûreté ou responsabilité pénale, et dans une perspective plus futuriste, aux potentielles conséquences d'une adoption globale des interfaces cerveau-machine et leurs implications quant à la définition de ce qu'est l'humain.

Ces réflexions sont rapportées dans le **Chapitre 8** de cette thèse.

VI. Conclusion

L'objectif de cette thèse était d'apporter de premières preuves de concept en vue du développement d'une interface cerveau-machine pour la restauration de la parole. En particulier, nous avons mis au point un système permettant de synthétiser de la parole intelligible à partir de trajectoires articulatoires, i.e. à partir de mouvements des articulateurs du conduit vocal comme la langue ou le palais mou. Ce synthétiseur a ensuite pu être contrôlé, en temps réel, par différents locuteur en condition de parole silencieuse, c'est-à-dire en articulant mais sans prononcer. Dans un second temps, nous avons développé une approche permettant de localiser et cartographier les aires corticales de la parole, pendant des chirurgies éveillées du cerveau, et directement sur le champ opératoire du neurochirurgien. Enfin, les signaux corticaux provenant des aires identifiées par cartographie ont pu être utilisés pour décoder, au-dessus du niveau de la chance, l'intention de parler du patient, aussi bien en parole orale qu'imaginée, mais aussi l'était de vibration des cordes vocales et, de manière moindre, les trajectoires articulatoires ou acoustiques de la parole.

Ces développements constituent un premier pas vers une interface cerveau-machine pour la restauration de la parole, et des travaux futurs devront confirmer ces résultats.