Understanding and Decoding Imagined Speech using Electrocorticographic Recordings in Humans

THÈSE N^{O} 7740 (2017)

PRÉSENTÉE LE 14 JUILLET 2017

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
CHAIRE FONDATION DEFITECH EN INTERFACE DE CERVEAU-MACHINE
PROGRAMME DOCTORAL EN NEUROSCIENCES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Stéphanie MARTIN

acceptée sur proposition du jury:

Prof. F. Schürmann, président du jury
Prof. J. D. R. Millán Ruiz, Prof. R. T. Knight, directeurs de thèse
Prof. A.-L. Giraud, rapporteuse
Prof. N. Ramsey, rapporteur
Prof. D. Van De Ville, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2017

« N'attendez pas que quelqu'un d'autre parle en votre nom. C'est vous qui pouvez changer le monde. » – Malala Yousafzai

Acknowledgements

I express my sincere gratitude to everyone that made the achievement of this thesis possible.

In particular, I would like to thank my parents, **Monika** and **Pierre-Yves**, and my brother **David** for supporting me throughout writing this thesis and my life in general. I am the person who I am today thanks to them.

Very special thanks go to **Giovanni**, who provided me with unfailing support and continuous encouragement through the process of writing this thesis. Thanks to him for being patient and comprehensive. Then, I also thank my friends – in particular **Amy, Charlotte and Marisa,** who have been my best friends for so many years now, and with whom I have shared so many giggles and adventures. In addition, thanks to **Luca, Cédric, Dave, Blaise, Nicky, Marion, Madelaine, Bastian, Yohan, Dimitri, Bernard, Isabelle, Jean** and many more for reminding me that there is a great world spinning out there. This accomplishment would not have been possible without them.

I am grateful to the members of my labs at EPFL and at UC Berkeley, for their aspiring guidance, invaluably constructive criticism and friendly advice during the project. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project. A particular acknowledgment goes to **Chris H.**, who guided me in my very first steps into this project, and gave me invaluable suggestions and ideas throughout this project. Similarly, a warm thank goes to **Iñaki** for the great collaborations and mentoring throughout the thesis. In particular, thanks to him for always being available at any time, and helping me out in my darkest moments of the writing. Thanks to **Michael, Ludovic, Arnaud, Luca, Chris S., Bastien, Sareh, Zahra** – to name just a few – for their positivity and enthusiasm. Thanks to them not only for being my colleagues, but friends as well.

I would like to acknowledge my supervisor Prof. **José Millán**, who gave me the opportunity to join his lab and his precious support, without which it would not have been possible to conduct this research. Thanks to him for making me discover the fundamentals of neuroengineering in the first place.

I express my sincere gratitude to my co-supervisor Prof. **Bob Knight** for the continuous support during my studies, for his motivation and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Thanks to him for trusting me, challenging me and giving me responsibilities.

Besides my advisors, my genuine thanks go to Dr. **Brian Pasley**, who shared his invaluable experience in the field. I thank him for being patient, critical and meticulous in decisive moments.

Last but not the least, I would like to thank the rest of my thesis committee, who kindly accepted to be part of this journey: Prof. **Anne-Lise Giraud**, Prof. **Dimitri Van De Ville**, Prof. **Nick Ramsey**, and Prof. **Felix Schürman**.

Abstract

Certain brain disorders, resulting from brainstem infarcts, traumatic brain injury, stroke and amyotrophic lateral sclerosis, limit verbal communication despite the patient being fully aware. People that cannot communicate due to neurological disorders would benefit from a system that can infer internal speech directly from brain signals. Investigating how the human cortex encodes imagined speech remains a difficult challenge, due to the lack of behavioral and observable measures. As a consequence, the fine temporal properties of speech cannot be synchronized precisely with brain signals during internal subjective experiences, like imagined speech. **This thesis aims at understanding and decoding the neural correlates of imagined speech (also called internal speech or covert speech), for targeting speech neuroprostheses.**

In this exploratory work, various imagined speech features, such as acoustic sound features, phonetic representations, and individual words were investigated and decoded from electrocorticographic signals recorded in epileptic patients in three different studies. This recording technique provides high spatiotemporal resolution, via electrodes placed beneath the skull, but without penetrating the cortex

In the first study, we reconstructed continuous spectrotemporal acoustic features from brain signals recorded during imagined speech using cross-condition linear regression. Using this technique, we showed that significant acoustic features of imagined speech could be reconstructed in seven patients.

In the second study, we decoded continuous phoneme sequences from brain signals recorded during imagined speech using hidden Markov models. This technique allowed incorporating a language model that defined phoneme transitions probabilities. In this preliminary study, decoding accuracy was significant across eight phonemes in one patients.

In the third study, we classified individual words from brain signals recorded during an imagined speech word repetition task, using support-vector machines. To account for temporal irregularities during speech production, we introduced a non-linear time alignment into the classification framework. Classification accuracy was significant across five patients.

In order to compare speech representations across conditions and integrate imagined speech into the general speech network, we investigated imagined speech in parallel with overt speech production and/or speech perception. Results shared across the three studies showed partial overlapping between imagined speech and speech perception/production in speech areas, such as superior temporal lobe, anterior frontal gyrus and sensorimotor cortex.

In an attempt to understanding higher-level cognitive processing of auditory processes, we also investigated the neural encoding of acoustic features during music imagery using linear regression. Despite this study was not directly related to speech representations, it provided a unique opportunity to quantitatively study features of inner subjective experiences, similar to speech imagery.

These studies demonstrated the potential of using predictive models for basic decoding of speech features. Despite low performance, results show the feasibility for direct decoding of natural speech. In this respect, we highlighted numerous challenges that were encountered, and suggested new avenues to improve performances.

Keywords

electrocorticography, imagined speech, neurobiology of language, neural predictive models, machine learning, linear regression, hidden Markov models, support-vector machine.

Résumé

Certaines personnes ne peuvent pas communiquer à cause de maladies ou traumatismes neurologiques, tel qu'un accident vasculaire cérébral, une sclérose amyotrophique latérale ou une lésion cérébrale. Ces personnes bénéficieraient d'un appareil médical qui puisse inférer le langage interne à partir de l'activité neuronale. **Cette thèse a pour but de comprendre et décoder les signaux cérébraux associés au langage interne, afin d'aider les personnes dans le besoin à communiquer.** Dans cette thèse exploratrice, nous avons étudié et décodé plusieurs caractéristiques différentes du langage interne, telles que les propriétés acoustiques, la représentation phonétique et la production de mots individuels. Pour cela, nous avons utilisé l'électrocorticographie, une technique d'enregistrement d'activé cérébrale dans laquelle les électrodes sont implantées dans la tête de patients épileptiques. Cette technique permet d'enregistrer l'activité cérébrale avec une grande précision spatiale et temporelle. En utilisant ce procédé, nous avons réalisé trois études.

Dans la première étude, nous avons reconstruit le spectrogramme du son à partir de l'activité cérébrale – un utilisant une régression linéaire. Dû au manque de mesure comportementale, nous ne pouvions pas régresser directement l'activité cérébrale au spectrogramme. Au lieu de cela, nous avons utilisé une autre stratégie basée sur le fait que parler à voix haute ou parler silencieusement ont des mécanismes neuronaux partiellement similaires. Le décodeur a été construit à base de données enregistrées lors de production de langage à voix haute, et ensuite, ce décodeur a été appliqué à des données enregistrées à voix silencieuse. En utilisant cette technique, nous avons démontrer que les caractéristiques acoustiques du langage peuvent être significativement reconstruite chez sept patients.

Dans la deuxième étude, nous avons décodé continuellement des séquences de phonèmes à partir de l'activité cérébrale – en utilisant le modèle de Markov caché (hidden Markov model). Cette technique permet d'intégrer un model linguistique basé sur les transitions phonémiques. Les performances du décodeur étaient significatives chez deux patients.

Dans la troisième étude, nous avons classifié des mots en utilisant des caractéristiques temporelles. Afin de prendre en compte les irrégularités dans la production de mots, nous avons introduit un alignement temporel dans le model de classification. Les performances étaient significatives chez cinq patients.

Dans le but de comparer ces représentations linguistiques du langage interne avec celles du langage oral, nous avons étudier le langage interne en parallèle avec la perception et la production oral de langage. Les résultats ont démontré des similarités dans diverses régions associées au langage, tel que le gyrus supérieur temporal, le gyrus frontal antérieur et le cortex sensorimoteur

Finalement, nous avons aussi étudié comment le cortex humain traite les caractéristiques acoustiques pendant la production interne de musique. Cette étude n'est pas directement liée au langage; toutefois, elle fournit une opportunité unique d'étudié de façon quantitatives les caractéristiques lors d'expériences subjectives, internes, tel qu'il est le cas lors production interne de langage.

Ces études représentent une preuve de concept pour le décodage de diverses caractéristiques du langage, et soulignent bon nombre de défi à relever pour concevoir un appareil médical réaliste.

Mots-clés

Électrocorticographie, langage interne, neurobiologie du langage, model prédictif, machine learning, régression linéaire, modèle de Markov caché, support-vector machine.

Content

List of figures

Chapter 1 Introduction

Disclaimer: This chapter is adapted from the following articles – with permissions of all co-authors and journals:

Martin S., Millán J.d.R., Knight R.T., and Pasley B.N. 2016. "The Use of Intracranial Recordings to Decode Human Language: Challenges and Opportunities." Brain and Language. doi:10.1016/j.bandl.2016.06.003.

My contribution: Conceptualization, visualization, writing – original draft preparation

Martin S., Brunner P., Holdgraf C., Heinze H.-J., Crone N.E., Rieger J., Schalk G., Knight R.T., Pasley B.N. 2014. "Decoding Spectrotemporal Features of Overt and Covert Speech from the Human Cortex." Frontiers in Neuroengineering. doi:10.3389/fneng.2014.00014.

My contribution: Conceptualization, formal analysis, methodology, visualization, writing – original draft preparation.

Holdgraf, C.R., Micheli, C., **Martin, S**., Knight, R.T., Rieger, J., Theunissen, F.E. . 2017. "Encoding and decoding models in cognitive electrophysiology." Frontiers in Systems Neuroscience (under revision).

My contribution: Conceptualization, visualization, writing – original draft preparation

Certain brain disorders limit verbal communication despite the patient being fully aware of what he wants to say. This neurological condition may results from brainstem infarcts, traumatic brain injury, stroke and amyotrophic lateral sclerosis, and affect more than two million people in the Unites States and far more around the world (Wolpaw et al. 2002). In order to help them to communicate, a few brain-computer interfaces have been proven useful (Farwell and Donchin 1988; Perdikis et al. 2014; Vansteensel et al. 2016; Pandarinath et al. 2017), but relied on indirect actions to convey information, such as performing mental tasks like a rotating cube, mental calculus or movements attempts. As an alternative, people with speech deficits would benefit from a communication system that can directly infer intended speech from brain signals – allowing them to interact more naturally with the world. **This thesis aimed at understanding and decoding the neural correlates associated with imagined speech, for targeting communication assistive technologies**. Imagined speech (also called inner speech, internal speech, silent speech, speech imagery, imagined speech or verbal thoughts) is defined here as the ability to generate internal speech representations, in the absence of any external speech stimulation or self-generated overt speech.

In this explorative work, we investigated the neural encoding mechanisms of various imagined speech representations, such as acoustic features in the early auditory system, phonemic features in intermediate levels of processing and individual words in later linguistic stages. For this, we used electrocorticographic (ECoG) recordings in the human cortex, a technique that allows monitoring brain activity with high spatial, temporal, and spectral resolution. We also evaluated the ability to decode these speech features for targeting communication devices. Each speech representation was decoded using specific machine-learning algorithms (e.g. regression, hidden Markov models and support-vector machines) – adapted for each task. Finally, in order to integrate imagined speech neural mechanisms in the general speech network, we investigated imagined speech in parallel with speech perception and speech production.

In this chapter, we introduce concepts that are general to all studies undertaken, whereas subsequent chapters develop aspects that are specific for each study. We first briefly describe the functional organization of speech – including current knowledge about imagined speech. Then, we present the properties of electrocorticography, together with its benefits to investigate human speech. We also describe neuro-computational modeling approaches used to decoding speech features, as well as the state-of-the-art in this field. Finally, we define precisely the objectives of this thesis.

1.1 Functional organization of speech

Speech is encoded in a widely distributed and complex network, whose activation depends on both linguistic context and brain region. It is well accepted that two main brain areas involved in speech comprehension and speech production are Wernicke's area (posterior superior and middle temporal gyrus/superior temporal sulcus) and Broca's area (posterior inferior frontal gyrus), respectively (*Figure 1-1*; see Price 2000; Démonet, Thierry, and Cardebat 2005; Hickok and Poeppel 2007 for reviews).

Speech comprehension involves multiple stages of neural representations in order to convert sound to meaning. The first stage in this process involves spectrotemporal analysis of the acoustic signal in early auditory cortices. This is followed by phonetic and phonological processing in the superior temporal lobe (Hickok and Poeppel 2007), in which continuous acoustic features are projected into categorical representations (Chang et al. 2010). Ultimately, higher levels of speech comprehension transform intermediate speech representations into conceptual and semantic representations in the so-called ventral stream (superior middle temporal lobe).

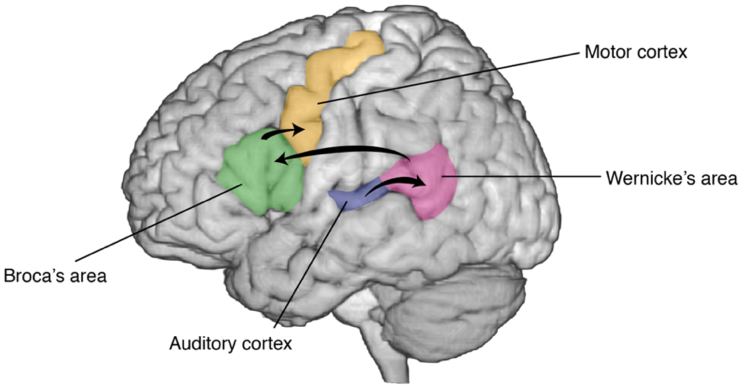


Figure 1-1 Speech network. *The major brain areas involved in speech processing are depicted (Price 2000). Early auditory cortices (Heschl's gyrus) involve spectrotemporal analysis of speech, and project into Wernicke's area (posterior superior temporal gyrus). The arcuate fasciculus connects Wernicke's area to Broca's area, which is involved in speech preparation and planning. Finally, the ventral sensorimotor cortex coordinates articulatory movements to produce audible speech.*

Alternatively, the dorsal stream (posterior dorsal temporal lobe, parietal operculum and posterior frontal lobe) is responsible for translating speech signals into articulatory representations. This network projects from primary auditory cortices to more dorsal aspects of the temporal lobe, and then to the posterior frontal lobe (Broca's area), as well as premotor and supplementary areas (Hickok and Poeppel 2007). Broca's area, which originally was recognized to be important for speech motor production, has recently been challenged regarding its role. Recent work suggested that Broca's area coordinates speech by mediating a cascade of activation from sensory representations to their corresponding articulatory gestures rather than actually providing the motor output (Flinker et al. 2015). Speech production itself is the result of coordinated and precise movements over rapid time scales in the ventral half of the lateral sensorimotor cortex (Levelt 1993; Gracco and Löfqvist 1994; Brown et al. 2009), where the articulators (i.e., lips, jaw, tongue and larynx) are organized somatotopically.

While actual speech perception and production have been extensively studied, the neural mechanisms underlying imagined speech remain poorly understood. Imagery-related brain activation could result from top-down induction mechanisms including memory retrieval (Kosslyn et al. 2001; Kosslyn 2005) and motor simulation (Tian et al. 2012; Guenther et al. 2006; Price 2011). In memory retrieval, perceptual experience may arise from stored information (objects, spatial properties and dynamics) acquired during actual speech perception and production experiences (Kosslyn 2005). In motor simulation, a copy of the motor cortex activity (efference copy) is forwarded to lower sensory cortices, enabling a comparison of actual with desired movement, and permitting online behavioral adjustments (Tian et al. 2012; Jeannerod 2003). Recently, functional magnetic resonance imaging studies have shown that imagined speech activates Wernicke's area (superior temporal gyrus; Yetkin et al. 1995; McGuire et al. 1996; Palmer et al. 2001; Shergill et al. 2001; Aleman 2004; Aziz-Zadeh et al. 2005; Geva, Correia, and Warburton 2011) and Broca's area (inferior frontal gyrus; Hinke et al. 1993; Huang, Carr, and Cao 2002) (see Price 2012; Perrone-Bertolotti et al. 2014 for reviews).

Although traditional brain imaging techniques have identified anatomical regions associated with speech, these methods lack the temporal resolution to investigate rapid temporal neural dynamics (Towle et al. 2008). In contrast, electrocorticography is a direct neural recording method that allows monitoring brain activity with high spatial, temporal, and spectral resolution (Ritaccio et al. 2014), and therefore is a good candidate to investigate speech processing.

1.2 Electrocorticographic recordings in the human brain

Electrocorticography (ECoG), also called intracranial recordings has been used for decades in patients with epilepsy to localize the seizure onset zone, prior to brain tissue resection. In such cases, clinical procedure requires temporary implantation of electrode grids or strips onto the cortical surface, either above (epidural) or below (subdural) the dura mater (*Figure 1-2*). In some cases, depth electrodes (stereoencephalography) are implanted in the cortex to identify epileptic sources functions (Halgren, Marinkovic, and Chauvel 1998). Because of its invasiveness, intracranial recordings are applied exclusively for clinical purposes; nevertheless, the implantation time provides a unique opportunity to investigate human brain functions. Electrode grids placed over the temporal cortex, frontal cortex and sensory motor cortex are the most relevant for investigating speech processes.

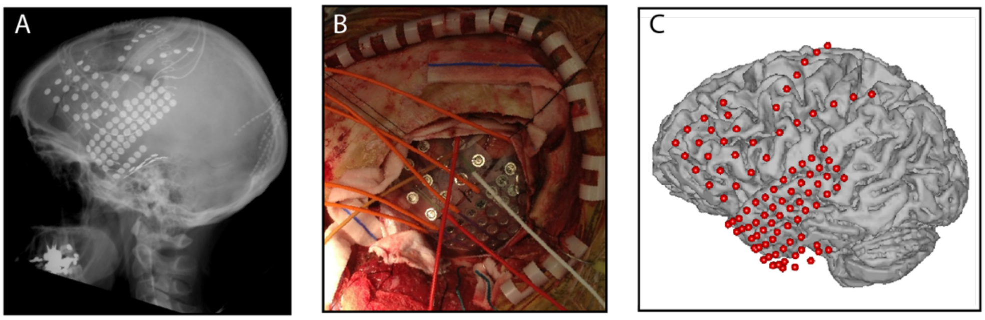


Figure 1-2 The ECoG grid and surgical placement. ***(A)*** *Radiography of electrode placement.* ***(B)*** *ECoG surgical placement.* ***(C)*** *Electrode positions in situ.*

ECoG has remarkable spatial (i.e., millimeter; Flinker et al. 2011) and spectral (0-500Hz; Staba et al. 2002) resolution, as well as higher amplitude (50-200μV) and signal-to-noise ratio – as compared to electroencephalography (EEG; centimeters, 0-40Hz, 10-20μV). The reduced frequency range in EEG is due to the 1/f rule resulting in amplitude reduction for higher frequencies. It is also less sensitive to artifacts such as those generated by the electrical activity from skeletal muscle movements (Ball et al. 2009). In addition, the electrodes cover broad brain areas compared to intracortical recordings. Although, scalp EEG has a better overall brain coverage (e.g. covers both hemispheres), it has increased distortion and smearing of the electrical signal through to the skull, and therefore a much lower spatial resolution. Finally, ECoG has much higher temporal resolution (millisecond) with respect to metabolic imaging techniques, such as functional magnetic resonance imaging and positron emission tomography (seconds).

Previous work has demonstrated the value of the ECoG signal in neuroprosthetic applications and for assistive technologies (Schalk et al. 2007; Felton et al. 2007). For example, ECoG recordings have been used in humans to control a one-dimensional (Leuthardt et al. 2004) and three-dimensional (Wang et al. 2013) cursor movements. Alternatively, formant frequencies (spectral peaks in the sound spectrum) were decoded in real time using intracortical recordings in the human motor cortex. The predicted speech was synthesized, and acoustically fed back to the user with a delay under 50ms (Guenther et al. 2009; Brumberg et al. 2010). Thus, ECoG represents a promising recording technique to investigate and decode neural correlates associated with human speech.

In particular, the high gamma frequency band (HG; 70-1200 Hz; also referred to as high frequency band) has been correlated with multiunit spike rate and asynchronous post-synaptic current of the underlying neuronal population (Manning et al. 2009; Lachaux et al. 2012; Buzsáki, Anastassiou, and Koch 2012). High gamma signal reliably tracks neural activity in many sensory modalities and correlates with cognitive functions, such as memory and speech (see Lachaux et al. 2012 for a review). The high gamma band has been correlated with spectrotemporal acoustic properties of speech in the superior temporal gyrus (Pasley et al. 2012; Kubanek et al. 2013), phonetic features (Chang et al. 2010), and articulatory features in the sensorimotor cortex (Bouchard et al. 2013). These studies demonstrate that many aspects of speech are robustly encoded in the high gamma frequency range of the ECoG signal.

Substantial efforts have aimed to develop new tools for analyzing these brain signals given the increasing amount of data recorded with intracranial electrode grids. The next section briefly introduces neural predictive models, which includes both encoding and decoding models. These approaches have been used as a quantitative method to test hypotheses about neural representation under study.

1.3 Predictive models in cognitive electrophysiology

Traditionally, cognitive processes have been investigated using a minimum set of stimulus type, that typically differ along a single dimension of interest (e.g. attended versus not attended target). Brain activity evoked by these two sets of stimuli are then averaged and compared in order to provide new insights about the neural mechanisms under study. While much has been learned about perception using these methods, they often rely on un realistic, natural scenarios, and do not allow reliable trial analysis. Predictive modeling – a relatively new alternative approach – allows researchers to apply multivariate neural features to rich, complex and naturalistic stimuli in the world. For instance, a sensory stimulus, a movement, or a cognitive state generates an electrophysiological brain response that is specific in time, frequency band and location. Neural encoding models attempt to characterize how these parameters are represented in the brain, and how they co-vary with changing patterns in the world. This type of predictive model essentially asks the question, given a particular stimulus or behavior representation, can we predict the resulting brain response. Conversely, decoding models attempt to predict information about conditions in the world that evoked a particular pattern of brain activity.

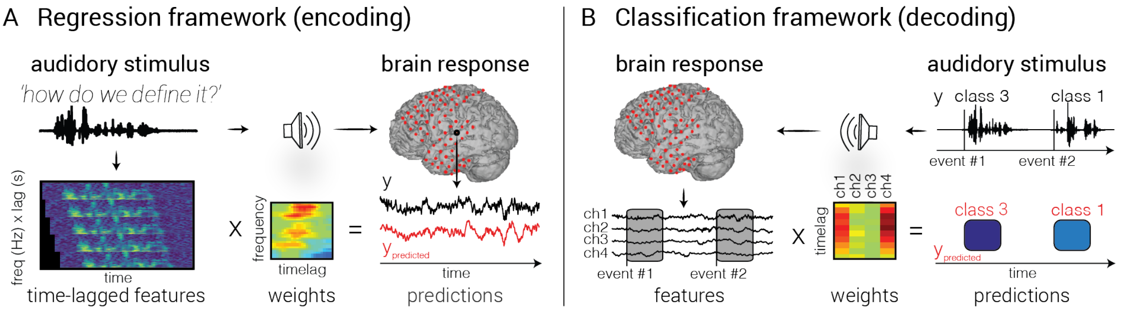


***Figure 1-3 Predictive models. (A)** In encoding models, the neural activity is predicted from the auditory representation by finding a set of electrode that minimizes the difference between true (black) and predicted (red) values, whereas in decoding models input and output features are reversed. **(B)** Discriminant features can also be used to predict a category among a finite number of choices.*

The simplest form of a predictive model uses a regression framework to link brain activity and a stimulus or mental state representation. For instance, in the case of encoding models, the neural activity at a given time is modeled as a weighted sum of stimulus features (*Figure 1-3 A*), whereas input/output are reversed in the case of decoding models. Classification is a special case of decoding models, in which the neural activity is identified as belonging to a discrete event type from a finite set of choices (*Figure 1-3 B*). Both types of predictive models can use various mathematical algorithms, ranging from simple regression techniques, to more complex non-linear approaches, such as hidden Markov models, support-vector algorithms and neural networks. Although the underlying math is different – and further developed in specific chapters – the general framework is common to all and consists in the following steps:

1. Feature extraction: in encoding models, input and output features are extracted from the stimulus and from the neural activity, respectively. In decoding models, input and output features are reversed. Examples of speech representations typically used in predictive models are the audio frequencies, the modulation rates, or phonemes for natural audio (*Figure 1-4*). For neural representations, amplitudes in specific frequency bands are typically extracted from the recorded electrophysiological signal.

2. Model estimation: models are estimated by mapping input features to output features. The weights are calculated by minimizing a metric of error between the predicted and actual output on a training set. Generally, these models are linear mapping, in which the output is a weighted sum of input features.

3. Validation: Once a model is fit, it is then validated on new unseen data – not used for training. To evaluate the accuracy, the predicted output is compared directly to the original representation.

***4.* Interpretation**: model weights reveal the relationship between brain activity and stimulus features. In the case of encoding model, the weights – also called receptive fields – identify the stimulus features encoded by a neuron or population of neurons.

Both encoding and decoding models are complementary, and have several benefits over traditional stimulus-contrast approaches. This included the ability to make predictions about new datasets, to take a multivariate approach to fitting model weights, and to use multiple feature representations with a complex, natural stimulus set. Encoding models are useful for exploring multiple levels of abstraction within a complex stimulus, and investigating how each affects activity in the brain. Alternatively, decoding models can be used to operate neuroprosthetic devices activity from patterns of brain activity, in order to help disabled individuals interacting with the world. For instance, decoding models have been used to move a cursor on the screen (Wolpaw et al. 1991), control a wheelchair (Millán et al. 2009), and predict 3D trajectories of a robotic arm (Hochberg et al. 2012).

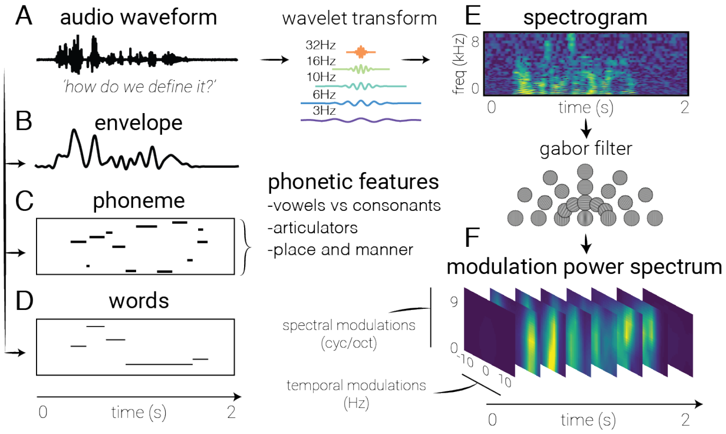


Figure 1-4 Feature extraction. *Several common auditory representations are shown for the same natural speech utterance (A). For instance, the raw audio waveform (A) can be low-pass filtered and squared to extract the speech envelope (B). Alternatively, a frequency decomposition of the raw auditory waveform can be used to generate a spectrogram that reflects spectrotemporal fluctuations over time (E). A two-dimensional Gabor decomposition of the spectrogram itself can be used to create the Modulation Power Spectrum of the stimulus (F). Alternatively, one may code the auditory stimulus with a categorical variable corresponding to linguistic features such as the presence of phonemes (C) or words (D). These approaches may be used to investigate the brain's response to higher-order features.*

In the next section, we first review ECoG studies that have employed predictive models to understand and decode cognitive states associated with various speech representations.

1.4 State-of-the-art research in speech modeling

Speech processing includes various processing steps – such as acoustic processing in the early auditory system, phonetic and categorical encoding in intermediate levels of processing and semantic and higher level of linguistic processes in later stages. In this section, we describe the various speech representations that were investigated using predictive models and ECoG signals recorded during speech perception and production. We also devote a specific section to the state-of-the-art in modeling imagined speech.

1.4.1 Early auditory speech representations

A well-established role of the early auditory system is to decompose speech and complex sounds into their elementary time-frequency components (Aertsen, Olders, and Johannesma 1981; Eggermont, Aertsen, and Johannesma 1983; Tian 2004). These spectrotemporal varying features, such as the envelope (*Figure 1-4 B*), the spectrogram (*Figure 1-4 E*), formant frequencies[1], and time-frequency modulations (*Figure 1-4 F*) carry essential phonological information. For instance, the speech envelope contains key information about speech intensity, rhythmic cadence and phonetic content that is primordial for understanding fluent, conversational speech. Alternatively, low and intermediate temporal modulations in the spectrogram (<4Hz) are linked with syllable rate, whereas fast modulations (>16Hz) are related to syllable onsets and offsets. Similarly, broad spectral modulations are associated with vowel formants, whereas narrow spectral modulations are associated with harmonics (Shamma 2003). These features are robust under a variety of noise conditions, and reflect important properties of speech intelligibility (Chi et al. 1999; Elliott and Theunissen 2009).

Recent studies have shown that spectrotemporal features could be reconstructed from high gamma frequency band using regression models. In particular, high frequency bands have been shown to reliably track the speech envelope (Figure 1-4 B) throughout the auditory system, and especially in early auditory cortices (Kubanek et al. 2013). Similarly, mel-frequency cepstral coefficients (Chakrabarti et al. 2013) and a spectrogram and a modulation-based representation (Pasley et al. 2012) were successfully reconstructed from high gamma signals. In a single case study, the recorded brain signals successfully predicted real-time formant frequencies using intracortical depth electrodes in the human motor cortex (Brumberg et al. 2010; Guenther et al. 2009). The audio signal was synthetized from the reconstructed acoustic features and fed back aurally to the patient within 50ms. These studies shows that neural decoding approaches can predict acoustic features of speech, and highlight the potential of using invasive recording techniques for building a neural-based speech synthesizer.

1.4.2 Intermediate speech representations

Linguistic elements, such as phonemes[2] are extracted from a complex acoustic speech signal. For instance, realizations of the same speech instance are highly variable, both between speakers and within the same individual (Kaur and Garg 2012). Due to the lack of invariance, it is difficult to find constant relations between a linguistic element (e.g. phoneme; *Figure 1-4 C*) and its acoustic manifestation. As such, the brain requires the rapid and effortless extraction of meaningful and invariant phonetic information from a highly variable continuous acoustic signal. The superior temporal gyrus plays a key role in transforming acoustic sounds into discrete linguistic units during speech perception (Chang et al. 2010; Mesgarani et al. 2014). Similar to what happens in the superior temporal gyrus, phonetic features are organized along acoustic features in the ventral sensorimotor cortex during speech perception (Cheung et al. 2016). Alternatively, during speech production, phonetic features are organized somatotopically based on articulators (lips, tongue, larynx and jaw; Bouchard et al. 2013). This means that sounds that have similar acoustic properties, but require different movements to produce them (e.g. 'b' and 'd'), activate the motor cortex in similar ways during listening, but not during speaking.

From a decoding perspective, several studies have shown successful classification of individual speech units into different categories, such as vowels and consonants (Pei et al. 2011), phonemes (Brumberg et al. 2011; Mugler et al. 2014) and syllables (Blakely et al. 2008). Recently, studies have

[1] concentrations of acoustic energy around particular frequencies in the speech waveform.

[2] The phoneme is the smallest speech unit and represent the building blocks for more complex speech utterances, such as words and sentences.

shown the ability to decode continuous phoneme sequences from high gamma activity during speech perception (Moses et al. 2016) and overt speech production (Herff et al. 2015). These studies – inspired from the field of speech recognition – included probabilistic language models in the decoding framework, and allowed the identification of words and sentences.

1.4.3 Higher levels of speech representations

Acoustic cues and phonetic information are embedded in streams of natural speech, and are arranged sequentially to convey more complex linguistic and semantic meanings (Vitevitch et al. 1999). The integration of elementary speech units into semantic objects has been found in the inferior frontal gyrus and superior temporal gyrus (Wang et al. 2011). In addition to bottom up processing streams, the neural activity associated with speech is modulated by top-down influences to help recognizing speech sounds. Various higher order neural mechanisms, such as expectations (Leonard et al. 2015; Leonard et al. 2016; Holdgraf et al. 2016), feedback monitoring loops (Chang et al. 2013; Houde and Chang 2015) attentional resources (Fritz et al. 2007; Mesgarani and Chang 2012) and lexical retrieval (Cibelli et al. 2015), allow natural verbal communication in noisy environments and background speech.

Interestingly, the neural activity of a listener that perceive a specific speech sound that has been replaced or masked by noise is grounded in acoustic representations. This suggest that even in the absence of a given speech sound, the neural patterns correlates with those that would have been elicited by the actual speech sound. This phenomenon is deeply influenced by internal forward processes related to prediction and integration of contextual knowledge (Leonard et al. 2016; Holdgraf et al. 2016). Similarly, behavioral studies have shown that people learn the statistical properties of sequentially arranged units, although they are not consciously aware of the probability distributions (Winkler, Denham, and Nelken 2009). For instance, hearing the sound /k/ followed by /uw/ ("koo") is more common than hearing /k/ followed by /iy/ ("kee"). A recent study showed that the neural response is modulated according to the language-level probability of surrounding sounds (Leonard et al. 2015).

1.4.4 Imagined speech representations

Mental imagery reproduces experiences and brain patterns similar to actual perception and production. For instance, behavioral studies have shown that structural and temporal properties of auditory features (Halpern 1989a; Pitt and Crowder 1992; Halpern et al. 2004; Intons-Peterson 1980; Halpern 1988) are preserved during auditory imagery. Brain imaging studies have demonstrated the activation of overlapping brain areas during speech imagery and perception (Kraemer et al. 2005; Yoo, Lee, and Choi 2001; Zatorre et al. 1996) and production (Palmer et al. 2001; Yetkin et al. 1995; Rosen et al. 2000; McGuire et al. 1996; Aleman 2004; Hinke et al. 1993). Despite these studies, it remains unclear what precise features – from early auditory to higher-level speech representations – are encoded during imagined speech.

Very few studies have exploited the advantageous properties of intracranial recordings to characterize imagined speech features, and none has employed encoding models to elucidate neural mechanisms during such cognitive processes. A recent study described the spatiotemporal evolution of high gamma activity during an overt and imagined word repetition using grand average techniques (Pei, et al. 2011; Leuthardt et al. 2012). In particular, they revealed high gamma changes in superior temporal lobe and the supramarginal gyrus during imagined speech repetition. Alternatively, decoding models have been used to classify neural activity associated with imagined speech features into categorical representations, i.e., covertly articulated isolated vowels (Ikeda et al. 2014), vowels and consonants during covert word production (Pei et al. 2011) and intended phonemes (Brumberg et al. 2011). In addition, a recent study compared the electrocorticographic activity related to overt versus covert conditions, and revealed a common network of brain regions (Brumberg et al. 2016).

The lack of physiological studies on imagined speech emphasizes the difficulties when investigating imagined speech representations. A reason for this is that the fine spectrotemporal fluctuation property of speech cannot be monitored precisely during imagined speech, due to the lack of behavioral output. Indeed, there are no verifiable or observable measures during internal subjective experiences like imagined speech. Therefore, it is difficult to time-lock brain activity to a measurable stimulus or behavioral state, and build predictive models that directly regress the neural activity to any speech representation. In addition, speech is subject to temporal irregularities (stretching/compressing, onset/offset delays) across repetitions. For instance, a predictive model that assumes fixed time features may not recognize two trials as belonging to the same class if the neural patterns are not temporally aligned. As a results, this leads to problems in exploiting the temporal resolution of electrocorticography, and to investigate what factor is encoded during imagined speech.

1.5 Objectives and main results

This thesis was an initial work aimed at better understanding imagined speech using electrocorticographic neural signals recorded in epileptic patients. We investigated various speech representations, such as acoustic sound features, phonemic features, and individual words. We also evaluated the ability to decode these speech features for targeting communication devices. For this, four different studies have been performed:

In **Chapter 2**, we reconstructed continuous acoustic features from high gamma neural activity recorded during imagined speech (Martin et al. 2014). Due to the lack of any measurable behavioral output, it is difficult to build a regression model that directly maps the neural activity to any behavioral metric or speech representation. Instead, given that speaking out loud and speaking in its mind share common neural mechanisms, we built the decoding model from an overt speech condition, and applied this decoder in the imagined speech. Using this technique, we showed that significant acoustic features of imagined speech could be reconstructed in seven patients. These findings also provided evidence that overt and imagined speech share underlying neural mechanisms.

In **Chapter 3**, we decoded continuous phoneme sequences from the high gamma neural activity recorded during imagined speech (Martin et al. 2017, in preparation). Until now, isolated phonemes were successfully decoded during imagined speech (Ikeda et al. 2014; (Pei, Barbour, et al. 2011) Brumberg et al. 2011), but these study failed to decode phoneme sequences during continuous speech. In order to label intended phonemes more accurately during imagined speech, we designed a karaoke-like task, in which visual words scrolling on the screen were divided into their phonemic representations. Then, we implemented a hidden Markov model – a modeling technique widely used in the field of speech recognition – that allows incorporating a language model. Using this approach, we showed that decoding accuracy was significant across eight phonemes in one patient. Although, preliminary results were promising, these findings need to be extended to a larger pool of participants, in order to draw conclusions.

In **Chapter 4**, we classified individual words from high gamma neural features recorded during an imagined speech word repetition task (Martin et al. 2016). Although words have been decoded during overt speech (Blakely et al. 2008), only phonemes were successfully predicted during imagined speech (Ikeda et al. 2014; Pei et al. 2011; Brumberg et al. 2011). To this end, we took advantage of the high temporal resolution offered by ECoG, and classified neural features in the time domain using support-vector machine. In order to account for temporal irregularities across trials, we introduced a non-linear time alignment into the classification framework. Results showed that the classification accuracy was significant across five patients. This study represents a proof of concept for basic decoding of speech imagery, yet the results to date are not yet robust enough for a clinical communication device.

In these studies, we investigated imagined speech in parallel with overt speech production and/or speech perception. This allowed comparing speech representations across conditions, and integrate imagined speech into the general speech network. Results revealed complex patterns of brain activity, that where partially overlapping across conditions. The most informative areas to decode speech were located in the superior temporal gyrus, inferior frontal gyrus and sensorimotor cortex, areas commonly associated with speech.

In **Chapter 5**, we investigated the neural encoding of acoustic features during music imagery using linear regression (Martin et al. 2017, under revision at Cerebral Cortex). This study was not directly related to speech representations, yet it helped understanding features of inner subjective experiences, such as auditory and speech imagery. This study relied on an extremely rare clinical case in which a patient undergoing neurosurgery for epilepsy treatment was also an adept piano player. To address this challenge, we recorded his electrocorticographic neural signals in a novel task design that permitted robust tracking of the spectrotemporal content of the intended music imagery. This experiment provided a unique opportunity to apply receptive field modeling techniques to quantitatively study neural encoding during music imagery. We found robust similarities between perception and imagery – in frequency and temporal tuning properties in auditory areas.

These studies represent a proof of concept for basic decoding of imagined speech, and delineate a number of key difficulties to usage of speech imagery neural representations for clinical applications. Accordingly, we emphasize numerous challenges that were encountered, and suggest new avenues that will hopefully help getting closer to a neural-based speech interface.

Chapter 2 Decoding spectrotemporal acoustic features of imagined speech

Disclaimer: This chapter is adapted from the following article – with permissions of all co-authors and journals:

Martin S., Brunner P., Holdgraf C., Heinze H.-J., Crone N.E., Rieger J., Schalk G., Knight R.T., Pasley B.N. 2014. "Decoding Spectrotemporal Features of Overt and Covert Speech from the Human Cortex." Frontiers in Neuroengineering. doi:10.3389/fneng.2014.00014.

My contribution: Conceptualization, formal analysis, methodology, visualization, writing – original draft preparation

2.1 Abstract

Decoding imagined speech is complicated by the lack of any measurable behavioral or acoustic output that is synchronized to brain activity. Thus, it is difficult to build a decoding model that directly regress the neural activity to any behavioral metric or speech representation. To counter this issue, we proposed an alternative strategy, based on the fact that auditory perception and auditory imagery activate overlapping brain regions. We built the decoding model on high gamma neural activity (70-150 Hz), and reconstructed spectrotemporal auditory features of self-generated overt speech. Then, the same decoding model was applied to reconstruct auditory speech features in the imagined speech condition. To evaluate performances, the reconstruction in the imagined speech condition was compared to the representation of the corresponding original sound spoken out loud – using a temporal realignment algorithm. Results showed that significant acoustic features of imagined speech could be reconstructed in five out of seven patients. This provided evidence that overt and imagined speech share underlying neural mechanisms, and the relationship between both depended on anatomy. The superior temporal gyrus, pre- and post-central gyrus provided the highest reconstruction information. In addition, the ability to decode acoustic features from imagined speech may provide a basis for development of a brain-based communication method for patients with disabling neurological conditions.

2.2 Introduction

Mental imagery produces experiences and neural activation patterns similar to actual perception. For instance, thinking of moving a limb activates the motor cortex, internal object visualization activates the visual cortex, with similar effects observed for each sensory modality (Kosslyn, Ganis, and Thompson 2001; Roth et al. 1996; Stevenson and Case 2005). Behavioral and neural studies have suggested that structural and temporal properties of auditory features, such as pitch (Halpern 1989a), timbre (Pitt and Crowder 1992; Halpern et al. 2004), loudness (Intons-Peterson 1980) and rhythm (Halpern 1988) are preserved during music imagery (Hubbard 2013). Less is known about the neural substrate of speech imagery, and how it compares to overt speech production. In this study, we investigated neural mechanisms associated with acoustic representations during overt and imagined speech, and evaluated the ability to reconstruct acoustic features from imagined speech data.

Increasing evidence suggests that speech imagery and perception activate the same cortical areas. Functional imaging studies have reported overlapping cortical regions during overt and imagined speech generation in inferior frontal lobes, sensorimotor cortex regions, supplementary motor areas, and anterior cingulate gyri (Palmer et al. 2001; Yetkin et al. 1995; Rosen et al. 2000). Transcranial magnetic stimulation over motor sites and inferior frontal gyrus induced speech arrest in both overt and imagined speech production (Aziz-Zadeh et al. 2005). Finally, brain lesion studies have shown high correlation between overt and imagined speech abilities, such as rhymes and homophones judgment (Geva, Bennett, et al. 2011) for patients with aphasia.

Despite findings of overlapping brain activation during overt and imagined speech (Palmer et al. 2001; Yetkin et al. 1995; Rosen et al. 2000; Aziz-Zadeh et al. 2005; McGuire et al. 1996; Aleman 2004; Hinke et al. 1993; Geva, Correia, et al. 2011), it is likely that imagined speech is not simply overt speech without moving the articulatory apparatus. Behavioral judgment studies showed that aphasic patients indicated inner speech impairment, while maintaining relatively intact overt speech abilities, while others manifested the reverse pattern (Geva, Bennett, et al. 2011). Similarly, imaging techniques showed different patterns of cortical activation during imagined compared to overt speech, namely in the premotor cortex, left primary motor cortex, left insula, and left superior temporal gyrus (Huang et al. 2002; Shuster et al. 2005; Pei et al. 2011). This suggests that brain activation maps associated with both tasks are dissociated at least in some cases (Geva et al. 2011; Aleman 2004; Shuster et al. 2005; Feinberg et al. 1986). The extent to which auditory perception and imagery engage similar underlying neural representations remains poorly understood.

To compare similarities between the neural representations of overt and imagined speech, and investigate if acoustic features of imagined speech can be decoded from ECoG signals, we employed neural decoding models to predict auditory features from brain activity – a technique that was already successfully applied to speech perception (Pasley et al. 2012). We hypothesized that speech perception and imagery share a partially overlapping neural representation in auditory cortical areas. We reasoned that if speech imagery and perception share neural substrates, the two conditions should engage similar neural representations. Thus, a neural decoding model trained from overt speech should be able to predict speech features in the imagined condition. In this study, we used a high gamma (70-150Hz) based neural decoding model, trained on continuous overt speech data. This model was then used to decode spectrotemporal auditory features from brain activity measured during an imagined speech condition. We showed that significant acoustic features of imagined speech could be reconstructed from models that were built from overt speech data. These findings supported shared neural mechanisms between both tasks, and provided a basis for development of a brain-based communication method for patients with disabling neurological conditions.

2.3 Material and Methods

2.3.1 Subjects and data acquisition

Electrocorticographic (ECoG) recordings were obtained using subdural electrode arrays implanted in 7 patients undergoing neurosurgical procedures for epilepsy. All patients volunteered and gave their informed consent (approved by the Albany Medical College Institutional Review Board) before testing. The implanted electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) consisted of platinum–iridium electrodes (4 mm in diameter, 2.3 mm exposed) that were embedded in silicon and spaced at an inter-electrode distance of 0.6-1cm. Grid placement and duration of ECoG monitoring were based solely on the requirements of the clinical evaluation (*Figure 2-1* ***Electrode location.***).

ECoG signals were recorded at the bedside using seven 16-channel g.USBamp biosignal acquisition devices (g.tec, Graz, Austria) at a sampling rate of 9600 Hz. Electrode contacts distant from epileptic foci and areas of interest were used for reference and ground. Data acquisition and synchronization with the task presentation were accomplished using BCI2000 software (Schalk et al. 2004; Schalk 2010). All channels were subsequently downsampled to 1,000 Hz, corrected for DC shifts, and band pass filtered from 0.5 to 200 Hz. Notch filters at 60 Hz, 120 Hz and 180 Hz were used to remove electromagnetic noise. The time series were then visually inspected to remove the intervals containing ictal activity as well as channels that had excessive noise (including broadband electromagnetic noise from hospital equipment or poor contact with the cortical surface). Finally, electrodes were re-referenced to a common average. The high gamma frequency band (70-150 Hz) was extracted using the Hilbert transform.

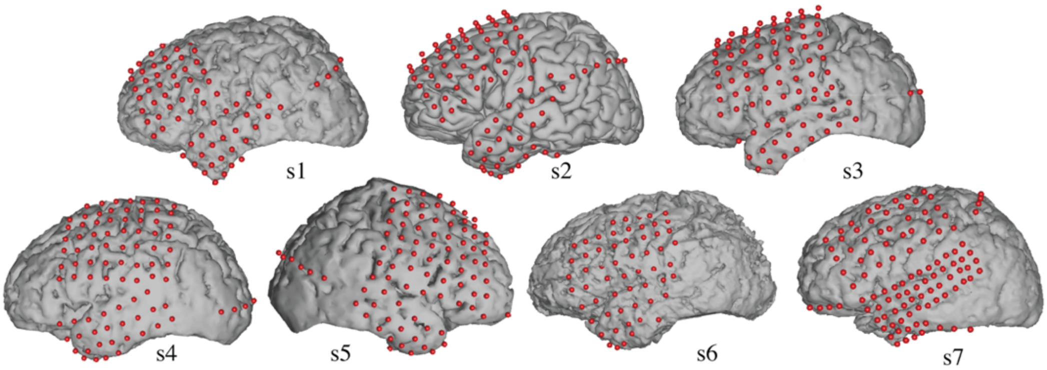


Figure 2-1 Electrode location. *Grid locations for each subject are overlaid on cortical surface reconstructions of each subject's MRI scan.*

In addition to the ECoG signals, we acquired the subject's voice through a dynamic microphone (Samson R21s) that was rated for voice recordings (bandwidth 80-12000 Hz, sensitivity 2.24 mV/Pa) and placed within 10 cm of the patient's face. We used a dedicated 16-channel g.USBamp to amplify and digitize the microphone signal in sync with the ECoG data. Finally, we verified the patient's compliance in the imagined task using an eye-tracker (Tobii T60, Tobii Sweden).

2.3.2 Experimental paradigms

The recording session included three conditions. In the first condition, text excerpts from historical political speeches or a children's story (i.e., Gettysburg Address (Roy and Basler 1955), JFK's Inaugural Address (Kennedy 1961), or Humpty Dumpy ("Mother Goose's Nursery Rhymes" 1867) were visually displayed on the screen moving from right to left at the vertical center of the screen. The rate of scrolling text ranged between 42-76 words/min, and was adjusted based on the subject's attentiveness, cognitive/verbal ability, and comfort prior to experimental recordings. In the first condition, the subject was instructed to read the text aloud (overt condition). In the second condition, the same text was displayed at the same scrolling rate, but the subject was instructed to read it silently (imagined condition). The third condition served as the control and was obtained while the subject was in a resting state condition (baseline control). For each condition, a run lasted between 6 and 8 min, and was repeated 2-3 times depending on the mental and physical condition of the subjects.

2.3.3 Auditory speech representations

We evaluated the predictive power of a neural decoding model to reconstruct two auditory feature representations: a spectrogram-based and a modulation-based representation. The spectrogram is a time-varying representation of the amplitude envelope at each acoustic frequency. This representation was generated by an affine wavelet transform of the sound pressure waveform using a 128 channel-auditory filter bank mimicking the frequency analysis of the auditory periphery (Chi, Ru, and Shamma 2005). The 128 acoustic frequencies of the initial spectrograms were subsequently downsampled to 32 acoustic frequency bins – with logarithmically spaced center frequencies ranging from 180-7,000 Hz.

The modulation representation is based on a non-linear transformation of the spectrogram. Spectral and temporal fluctuations reflect important properties of speech intelligibility. For instance, comprehension is impaired when temporal modulations (<12 Hz) or spectral modulations (4 cycles/kHz) are removed (Elliott and Theunissen 2009). In addition, low and intermediate temporal modulation rates (< 4 Hz) are linked with syllable rate, whereas fast modulations (> 16 Hz) are related to syllable onsets and offsets. Similarly, broad spectral modulations are associated with vowel formants, whereas narrow spectral modulations are associated with harmonics (Shamma 2003). The modulation representation was generated by a 2-D affine wavelet transform of the 128 channel auditory spectrogram. The bank of modulation-selective filters spanned a range of spectral scales (0.5–8 cycle/octave) and temporal rates (1–32 Hz), and was estimated from studies of the primary auditory cortex (Chi et al. 1999). The modulation representation was obtained by taking the magnitude of the complex-valued output of the filter bank, and subsequently reduced to 60 modulation features (5 scales x 12 rates) by averaging along the frequency dimension. These operations were computed using the NSL Matlab toolbox (http://www.isr.umd.edu/Labs/NSL/Software.htm). In summary, the neural decoding model predicted 32 spectral frequency features and 60 rate and scale features in the spectrogram-based and modulation-based speech representation, respectively.

2.3.4 Reconstruction procedure

Overt speech decoding

The decoding model was a linear mapping between neural activity and the speech representation (*Figure 2-2 A*). It modeled the speech representation (spectrogram or modulation) as a linear weighted sum of activity at each electrode as follows:

$$\hat{S}(t,p) = \sum_{\tau}\sum_{n} g(\tau,p,n)R(t-\tau,n),$$

Where $R(t-\tau,n)$ is the high gamma activity of electrode n at time $t-\tau$, where τ is the time lag ranging between -500ms and 500ms. $\hat{S}(t,p)$ is the estimated speech representation at time t and speech feature p, where p is one of 32 acoustic frequency features in the spectrogram-based representation (or one of 60 modulation features (5 scales x 12 rates) in the modulation-based representation. Finally, $g(\tau,p,n)$ is the linear transformation matrix, which depends on the time lag, speech feature, and electrode channel. Both speech representations and the neural high gamma response data were synchronized, downsampled to 100 Hz, and standardized to zero mean and unit standard deviation prior to model fitting.

Model parameters, the matrix *g* described above, were fit using gradient descent with early stopping regularization – an iterative linear regression algorithm. We used a jackknife resampling technique to fit separately between 4 and 7 models (Efron 1982), and then averaged the parameter estimates to yield the final model. To deal with auto-correlated neural activity and speech features, the data were first divided into 7-seconds blocks. Then, 90% of the data were randomly partitioned into training set and 10% into testing set. Within the training set, 10% of the data were used to monitor out-of-sample prediction accuracy to determine the early stopping criterion and minimize overfitting. The algorithm was terminated after a series of 30 iterations failing to improve performance. Finally, model prediction accuracy was evaluated on the independent testing set. Model fitting was performed using the STRFLab MATLAB toolbox (http://strflab.berkeley.edu/).

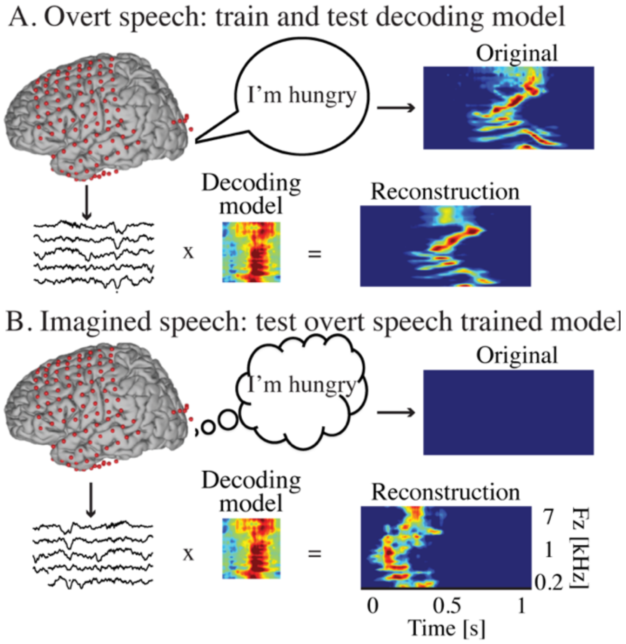


Figure 2-2 Reconstruction framework. ***(A)*** *The overt speech condition was used to train and test the accuracy of a neural-based decoding model to reconstruct spectrotemporal features of speech. The reconstructed patterns were compared to the true original (spoken out loud) speech representation (spectrogram or modulation-based).* ***(B)*** *During imagined speech, there is no behavioral output, which prevents building a decoding model directly from imagined speech data. Instead, the decoding model trained from the overt speech condition is used to decode imagined speech neural activity. The imagined speech reconstructed patterns were compared to identical speech segments spoken aloud during the overt speech condition (using dynamic time warping realignment).*

Imagined speech decoding

Decoding imagined speech is complicated by the lack of any measurable behavioral or acoustic output that is synchronized to brain activity. In other words, there is no simple ground truth by which to evaluate the accuracy of the model when a well-defined output is unavailable. To address this, we used the following approach. First, the decoding model was trained using data from the overt speaking condition. Second, the same model was applied to data from the imagined condition to predict speech features imagined by the subject (*Figure 2-2 B*), as follows:

$$\hat{S}_{imagined}(t,p) = \sum_{\tau}\sum_{n} g(\tau,p,n) R_{covert}(t-\tau,n),$$

where $\hat{S}_{imagined}(t,p)$ is the predicted imagined speech representation at time t and speech feature p, and $R_{imagined}(t-\tau,n)$ is the high gamma neuronal response of electrode n at time $t-\tau$, where τ is the time lag ranging between -500ms and 500ms. Finally, $g(\tau,p,n)$ is the linear model trained from the overt speech condition. To evaluate prediction accuracy during imagined speech, we made the assumption that the imagined speech representation should match the spectrotemporal content of overt speech. In this sense, overt speech is used as the "ground truth". Because subjects read the same text segments in both overt and imagined conditions, we computed the similarity between the imagined reconstructions and the corresponding original speech sounds recorded during the overt condition. To account for timing differences between conditions, we used dynamic time warping (DTW) to realign the imagined reconstruction to the original overt speech sound, as described in the next section.

2.3.5 Dynamic time warping

We used a dynamic time warping algorithm to realign the imagined speech reconstruction with the corresponding spoken audio signal from the overt condition, allowing a direct estimate of the imagined reconstruction accuracy. For the overt speech reconstructions, dynamic time warping was not employed, unless otherwise stated. DTW is a standard algorithm used to align two sequences that may vary in time or speed (Sakoe and Chiba 1978; Giorgino 2009). The idea behind DTW is to find the optimal path through a local similarity matrix d, computed between every pair of elements in the query and template time series, $X \in R^{P\,x\,N}$and $Y \in R^{P\,x\,M}$ as follows:

$$\mathrm{d(n,m) = f(x_n, y_m)}, \qquad \mathrm{d} \in \mathbb{R}^{\mathrm{NxM}},$$

where d is the dissimilarity matrix at time n and m, f can be any distance metric between sequence x and y at time n and m, respectively. In this study, we used the Euclidean distance, defined as $\mathrm{d(n,m)} = \sqrt{\sum_p^P (x_{np} - y_{mp})^2}$. Given ϕ, the average accumulated distortion between both warped signals is defined by:

$$d_{\varphi}(x,y) = \sum_{k=1}^{K} \frac{d(\varphi_x(k), \varphi_y\,(k))}{C_{\varphi}},$$

where φ_x and φ_y are the warping functions of length K (that remap the time indices of X and Y, respectively), and C_{φ} is the corresponding normalization constant (in this case N+M), ensuring that the accumulated distortions are comparable along different paths. The optimal warping path ϕ, chooses the indices of $X$and Yin order to minimize the overall accumulated distance.

$$D(X,Y) = \min_{\varphi} d_{\varphi}(X,Y),$$

where D is the accumulated distance or global dissimilarity. The alignment was computed using Rabiner-Juan step patterns (type 3; Rabiner 1993). This step pattern constrained the sets of allowed transitions between matched pairs to:

$$[\varphi_x(k+1) - \varphi_x(k),\ \varphi_y(k+1) - \varphi_y(k)] \in \{(1,2),(2,1),(1,1)\}$$

In addition, we assumed that the temporal offsets between imagined speech and original overt speech would be less than 2 sec, and thus introduced a global constraint – the Sakoe-Chiba band window (Sakoe and Chiba 1978), defined as follows:

$$\left|\varphi_x(k) - \varphi_y(k)\right| \leq T$$

where T = 2 sec was the chosen value that defines the maximum-allowable width of the window. Finally, to reduce computational load, the entire time series was broken into 30 sec segments, and warping was applied on each individual pair of segments (overt, imagined, or baseline control reconstruction warped to original speech representation). The warped segments were concatenated and the reconstruction accuracy was defined on the full time series of warped data. The DTW package in R (Giorgino 2009) was used for all analyses.

Baseline control condition (resting state)

To assess statistical significance of the imagined reconstruction accuracy, we applied the same decoding steps to a baseline control condition taken from data recorded during a separate resting state recording session. The overt speech decoding model was applied to neural data from the baseline control, as follows:

$$\hat{S}_{baseline}(t,p) = \sum_{\tau}\sum_{n} g(\tau,p,n)R_{baseline}(t-\tau,n),$$

where $\hat{S}_{baseline}(t,p)$ is the predicted baseline reconstruction at time t and speech feature p, and $R_{baseline}(t-\tau,n)$ is the high gamma neural response during resting state. Finally, $g(\tau,p,n)$ is the linear model trained from the overt speech condition. We also used DTW to realign the baseline control reconstruction with the spoken audio signal from the overt condition, allowing a direct estimate of the control condition decoding predictions.

2.3.6 Evaluation

In the overt speech condition, reconstruction accuracy was quantified by computing the correlation coefficient (Pearson's r) between the reconstructed and original speech representation using data from the independent test set. For each cross-validation resample, we calculated one correlation coefficient for each speech feature over time – leading to 32 correlation coefficients (one for each acoustic frequency features) for the spectrogram-based model and 60 correlation coefficients (5 scale x 12 rate features) for the modulation-based model. Overall reconstruction accuracy was reported as the mean correlation over resamples and speech components (32 and 60 for the spectrogram and modulation representation, respectively). Standard error of the mean (SEM) was calculated by taking the standard deviation of the overall reconstruction accuracy across resamples. To assess statistical significance, overt speech reconstruction accuracy was compared to the accuracy obtained from the baseline control condition (resting state).

In the imagined speech condition, we first realigned the reconstructions and original overt speech representations using dynamic time warping. Then, we computed the overall reconstruction accuracy using the same procedure as in the overt speech condition. To evaluate statistical significance, DTW was also applied to the baseline control condition prior to assessing the overall reconstruction accuracy.

To further assess the predictive power of the reconstruction process, we evaluated the ability to identify specific blocks of speech utterances within the continuous recording. First, 24-140 segments of speech utterances (5 sec duration) were extracted from the original and reconstructed spectrogram representations. Second, a confusion matrix was constructed where each element contained the similarity score between the target reconstructed segment and the original reference segments from the overt speech spectrogram. To compute the similarity score between each target and reference segment, DTW was applied to temporally align each pair and the mean correlation coefficient was used as the similarity score. The confusion matrix reflects how well a given reconstructed segment matches its corresponding original segment versus other candidates. The similarity scores were sorted, and identification accuracy was quantified as the percentile smaller than the rank of the correct segment (Pasley et al. 2012). At chance level, the expected percentile rank is 0.5, while perfect identification is 1.0.

To define the most informative areas for overt speech decoding accuracy, we isolated for each electrode its corresponding decoding weights, and used the electrode-specific weights to generate a separate reconstruction for each electrode. This allowed calculating a reconstruction accuracy correlation coefficient for each individual electrode. We applied the same procedure to the baseline condition. Baseline reconstruction accuracy was subtracted from the overt values to generate subject-specific informative area maps (*Figure 2-6*). The same technique was used in the imagined speech condition, except that DTW was applied to realign separately each electrode-specific reconstruction to the original overt speech. Similarly, baseline reconstruction accuracy (with DTW realignment) was subtracted from the imagined values to define the informative areas (*Figure 2-9*).

2.3.7 Statistics

To assess statistical significance for the difference between overt speech and baseline control reconstruction accuracy, we used Hotelling's t statistic with a significance level of p<10-5. This test accounts for the dependence of the two correlations on the same group (i.e. both correlations are relative to the same original overt speech representation) (Hotelling 1940; Birk 2013). It evaluates whether the correlations between overt speech reconstruction accuracy and baseline reconstruction accuracy differed in magnitude taking into account their intercorrelation, as follows:

$$t = \frac{(r_{jk} - r_{jh})\sqrt{(n-3)(1+r_{kh})}}{\sqrt{2}|R|}$$

where r_{jk} is the correlation between original overt speech and reconstruction, r_{jh} is the correlation between original overt speech and baseline reconstruction and r_{kh} is the correlation between overt speech reconstruction and baseline reconstruction; df = n – 3 is the effective sample size (Kaneoke et al. 2012) and where

$$|R| = 1 + 2r_{jk}r_{jh}r_{kh-}r_{jk}{}^2 - r_{jh}{}^2 - r_{kh}{}^2$$

At the population level, statistical significance was performed using Student's t-tests (p<10-5) after first applying Fisher's Z transform to convert the correlation coefficients to a normal distribution (Fisher 1915).

Test of significance in the imagined speech condition was equivalent to the overt condition (p<0.05; Hotelling's t test), except that the reconstructions and original overt speech representations were first realigned using dynamic time warping. Since DTW induces an artificial increase in correlation by finding an optimal warping path between any two signals (including potential noise signals), this procedure causes the accuracy for baseline reconstruction to exceed zero correlation. However, because the equivalent data processing sequence was applied to both conditions, any statistical differences between the two conditions were due to differences in the neural input signals.

At the population level, we directly compared the reconstruction accuracy in all three conditions (overt, imagined and baseline control). DTW realignement to the original overt speech was first applied separately for each condition. Reconstruction accuracy was computed as the correlation between the respective realigned pairs. Statistical significance was performed using Fisher's Z transform and one-way ANOVA (p<10-6), followed by post hoc t-test (p<10-5 for overt speech; p<0.005 for imagined speech).

For individual subjects, significance of identification rank was computed using a randomization test ($p<10^{-5}$ for overt speech; p<0.005 for imagined speech; p>0.5 for baseline control). We shuffled the segment label in the candidate set 10,000 times to generate a null distribution of identification ranks under the hypothesis that there is no relationship between target and reference speech segments. Time-varying speech representations are auto-correlated. To maintain temporal correlations in the data, and preserve the exchangeability of the trial labels, the length of the extracted segments was chosen sufficiently longue (5 seconds). The proportion of shuffled ranks greater than the observed rank yields the p-value that the observed accuracy is due to chance. Identification accuracy was assessed for each of the three experimental conditions (overt reconstruction, imagined reconstruction, baseline control reconstruction). At the population level, significant identification performance was tested using a one-sided, one-sample t-test ($p<10^{-5}$ for overt speech; p<0.05 for imagined speech ; p>0.5 for baseline control).

For the informative electrode analysis, statistical significance of overt speech reconstruction was determined relative to the baseline condition using Hotelling's t statistic (Hotelling's t test). Electrodes were defined as "informative" if the overt speech reconstruction accuracy was significantly greater than baseline (p<0.05; Hotelling's t test with Bonferroni correction). The same procedure was used for imagined speech informative areas (p<0.05; Hotelling's t test with Bonferroni correction), except that DTW was used in both imagined speech and baseline control condition.

2.3.8 Coregistration

Each subject had postoperative anterior–posterior and lateral radiographs, as well as computer tomography (CT) scans to verify ECoG grid locations. Three-dimensional cortical models of individual subjects were generated using pre-operative structural magnetic resonance (MR) imaging. These MR images were co-registered with the post-operative CT images using Curry software (Compumedics, Charlotte, NC) to identify electrode locations. Electrode locations were assigned to Brodmann areas using the Talairach Daemon (http://www.talairach.org, (Lancaster et al. 2000). Activation maps computed across subjects were projected on this 3D brain model, and were generated using a custom Matlab program (Gunduz et al. 2012).

2.4 Results

2.4.1 Overt speech reconstruction

The overall spectrogram reconstruction accuracy for overt speech was significantly greater than baseline control reconstruction accuracy in all individual subjects ($p<10^{-5}$; Hotelling's t-test, *Figure 2-3 A*). At the population level, mean overall reconstruction accuracy averaged across all subjects (N = 7) was also significantly higher than baseline control condition (r=0.41, $p<10^{-5}$; Fisher's Z transform followed by paired two-sample t-test). The baseline control reconstruction accuracy was not significantly different from zero (r=0.0, p>0.1; one-sample t-test; dashed line; *Figure 2-3 A*). Group averaged reconstruction accuracy for individual acoustic frequencies ranged between r=~0.25 – 0.5 (*Figure 2-3 B*).

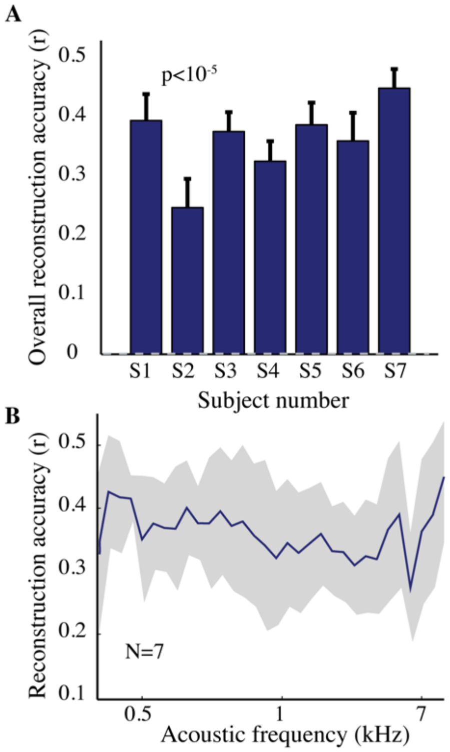


Figure 2-3 Overt speech reconstruction accuracy for the spectrogram-based speech representation. (A) *Overall reconstruction accuracy for each subject using the spectrogram-based speech representation. Error bars denote standard error of the mean (SEM). Overall accuracy is reported as the mean over all features (32 acoustic frequencies ranging from 0.2-7 kHz). The overall spectrogram reconstruction accuracy for the overt speech was greater than baseline control reconstruction accuracy in all individuals ($p<10^{-5}$; Hotelling's t-test). Baseline control reconstruction accuracy was not significantly different from zero (p>0.1; one-sample t-test; grey dashed line)* ***(B)*** *Reconstruction accuracy as a function of acoustic frequency averaged over all subjects (N=7) using the spectrogram model. Shaded region denotes SEM over subjects.*

An example of a continuous segment of the original and reconstructed spectrogram is depicted for a subject with left hemispheric coverage in *Figure 2-4 A*. In this subject, the reconstruction quality permitted accurate identification of individual decoded speech segments (*Figure 2-4 B*). The median identification rank (0.87, N=123 segments) was significantly greater than chance level (0.5, $p<10^{-5}$; randomization test). Identification performance was significant in each individual subject ($p<10^{-5}$; randomization test). Across all subjects, identification performance was significant for overt speech reconstruction ($rank_{overt}$=0.91 > 0.5, $p<10^{-6}$; one-sided one-sample t-test), whereas the baseline control condition was not significantly greater than chance level ($rank_{baseline}$ = 0.48 > 0.5, p>0.5 one-sided one-sample t-test).

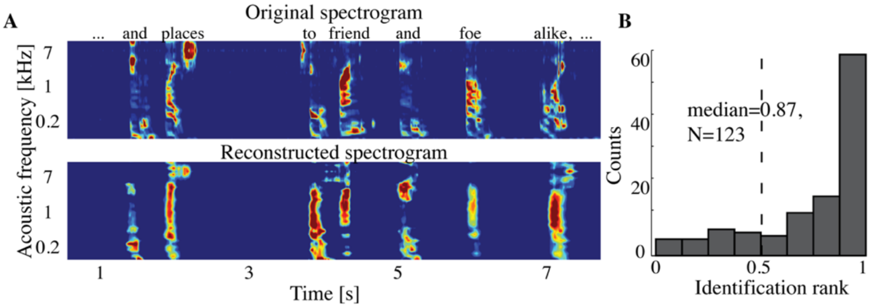


Figure 2-4 Overt speech reconstruction. (A) *Top panel: segment of the original sound spectrogram (subject's own voice), as well as the corresponding text above it. Bottom panel: same segment reconstructed with the decoding model.* ***(B)*** *Identification rank. Speech segments (5 sec) were extracted from the continuous spectrogram. For each extracted segment (N=123) a similarity score (correlation coefficient) was computed between the target reconstruction and each original spectrogram of the candidate set. The similarity scores were sorted and identification rank was quantified as the percentile rank of the correct segment. 1.0 indicates the target reconstruction matched the correct segment out of all candidate segments; 0.0 indicates the target was least similar to the correct segment among all other candidates; (dashed line indicates chance level = 0.5; median identification rank = 0.87; $p<10^{-5}$; randomization test).*

We next evaluated reconstruction accuracy of the modulation representation. The overall reconstruction accuracy was significant in all individual subjects ($p<10^{-5}$; Hotelling's t-test; *Figure 2-5 A*). a population level, mean overall reconstruction accuracy averaged over all patients (N = 7) was also significantly higher than the baseline reconstruction (r=0.55, $p<10^{-5}$; Fisher's Z transform followed by paired two-sample t-test). The baseline control reconstruction accuracy was not significantly different from zero (r=0.02, p>0.1; one-sample t-test; dashed line; *Figure 2-5 A*). Group averaged reconstruction accuracy for individual rate and scale was highest for temporal modulations above 2 Hz (*Figure 2-5 B*).

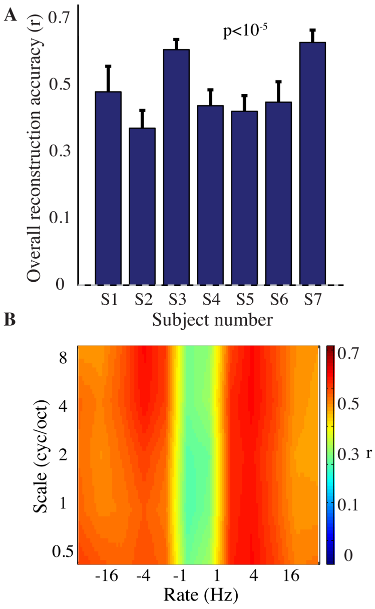


Figure 2-5 Overt speech reconstruction accuracy for the modulation-based speech representation. (A) *Overall reconstruction accuracy for each subject using the modulation-based speech representation. Error bars denote SEM. Overall accuracy is reported as the mean over all features (5 spectral and 12 temporal modulations ranging between 0.5-8 cyc/oct and -32-32 Hz, respectively). The overall modulation*

reconstruction accuracy for the overt speech was greater than baseline control reconstruction accuracy in all individuals ($p<10^{-5}$; Hotelling's t-test). Baseline control reconstruction accuracy was not significantly different from zero (p>0.1; one-sample t-test; grey dashed line). ***(B)*** *Reconstruction accuracy as a function of rate and scale averaged over all subjects (N=7).*

Figure 2-6 shows the significant informative areas (map thresholded at $p<0.05$; Bonferroni correction), quantified by the electrode-specific reconstruction accuracy. In both spectrogram and modulation-based representations the most accurate sites for overt speech decoding were localized to the superior temporal gyrus, pre and post central gyrus, consistent with previous spectrogram decoding studies (Pasley et al. 2012).

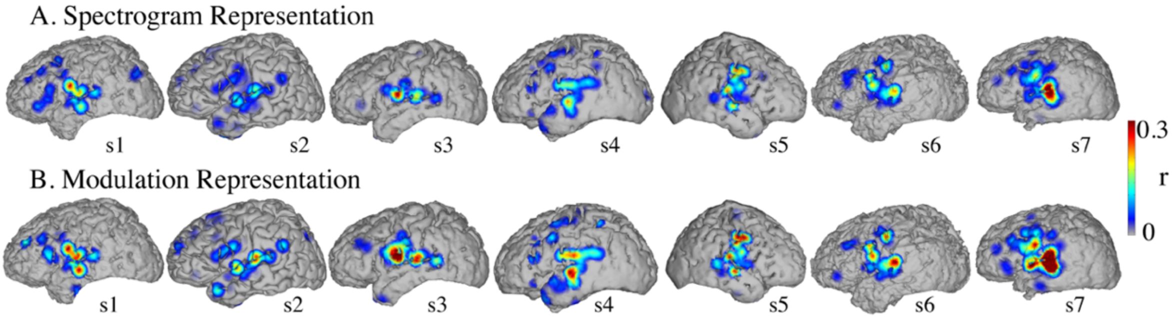


Figure 2-6 Overt speech predictive power*. Reconstruction accuracy correlation coefficients were computed separately for each individual electrode and for both overt and baseline control conditions (see section 3.1.3 for details). The plotted correlation values are calculated by subtracting the correlation during baseline control from the overt condition. The informative area map was thresholded to p<0.05 (Bonferroni correction)* ***(A)*** *Spectrogram-based reconstruction accuracy* ***(B)*** *modulation-based reconstruction accuracy.*

2.4.2 Imagined speech reconstruction

Figure 2-7 shows the overall reconstruction accuracy for overt speech, imagined speech, and baseline control after DTW realignment to the original overt speech was applied separately for each condition. The overall reconstruction accuracy for imagined speech was significantly higher than the control condition in 5 out of 7 individual subjects ($p<0.05$; Hotelling's t-test; $p>0.05$ for the non-significant subjects). At the population level, there was a significant difference in the overall reconstruction accuracy across the three conditions (overt, imagined and baseline control; $F_{(2,\ 18)} = 35.3$, $p<10^{-6}$; Fisher's Z transform followed by one-way ANOVA). Post-hoc t-tests confirmed that imagined speech reconstruction accuracy was significantly lower than overt speech reconstruction accuracy ($r_{imagined} = 0.34 < r_{overt} = 0.50$, $p<10^{-5}$; Fisher's Z transform followed by paired two-sample t-test), but higher than the baseline control condition ($r_{imagined} = 0.34 > r_{baseline} = 0.30$, $p<0.005$; Fisher's Z transform followed by a paired two-sample t-test).

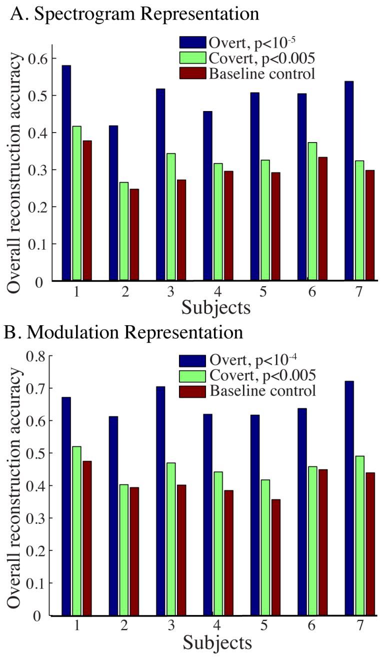


***Figure 2-7 Reconstruction accuracy using DTW realignment**. Overall reconstruction accuracy for each subject during overt speech, imagined speech and baseline control conditions after dynamic time warping realignment. **(A)** Spectrogram-based representation **(B)** Modulation-based representation.*

Figure 2-8 illustrates a segment of the reconstructed imagined speech spectrogram and its corresponding overt segment (realigned with DTW). We next evaluated identification performance (N=123 segments) for imagined speech and baseline control conditions in this subject (*Figure 2-8 B*). In the imagined speech condition, the median identification rank equaled 0.62, and was significantly higher than chance level of 0.5 (p<0.005; randomization test), whereas the baseline control condition was not significant (median identification rank = 0.47, p>0.5; randomization test). Several of the remaining subjects exhibited a trend toward higher identification performance, but were not significant at the p<0.05 level (*Figure 2-8*; randomization test). At the population level, mean identification performance across all subjects was significantly greater than chance for the imagined condition ($\text{rank}_{\text{imagined}}$ = 0.55 > 0.5, p<0.05; one-sided one-sample t-test), and not significant for the baseline control ($\text{rank}_{\text{baseline}}$ = 0.48 > 0.5, p>0.5; one-sided one-sample t-test). These results provide evidence that neural activity during auditory speech imagery can be used to decode spectrotemporal features of imagined speech.

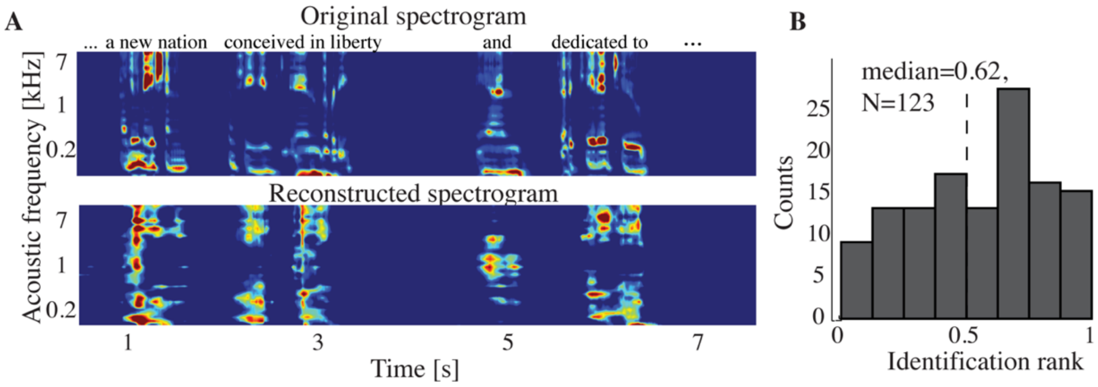


Figure 2-8 Imagined speech reconstruction (A) *Top panel: a segment of the overt (spoken out loud) spectrogram representation. Bottom panel: the same segment reconstructed from neural activity during the imagined condition using the decoding model.* ***(B)*** *Identification rank. Speech segments (5 sec) were extracted from the continuous spectrogram. For each target segment (N=123) a similarity score (correlation coefficient) was computed between the target reconstruction and each original spectrogram in the candidate set. The similarity scores were sorted and identification rank was quantified as the percentile rank of the correct segment. 1.0 indicates the target reconstruction matched the correct segment out of all candidate segments; 0.0 indicates the target was least similar to the correct segment among all other candidates. (dashed line indicates chance level = 0.5; median identification rank = 0.62; p<0.005; randomization test).*

Reconstruction accuracy for the modulation-based imagined speech condition was significant in 4 out of 7 individuals (p<0.05; Hotelling's t-test; p>0.1 for non-significant subjects; *Figure 2-7 B*). At the population level, the overall reconstruction accuracy across the three conditions (overt, imagined and baseline control) was significantly different ($F_{(2, 18)} = 62.1$, $p<10^{-6}$; one-way ANOVA). Post-hoc t-tests confirmed that imagined speech reconstruction accuracy was significantly lower than overt speech reconstruction accuracy ($r_{imagined} = 0.46 < r_{overt} = 0.66$, $p<10^{-5}$; Fisher's Z transform followed by a paired two-sample t-test), but higher than the baseline control condition ($r_{imagined} = 0.46 > r_{baseline} = 0.42$, p<0.005; Fisher's Z transform followed by a paired two-sample t-test).

Significant informative areas (map thresholded at p<0.05; Bonferroni correction), quantified by the electrode-specific reconstruction accuracy are shown in *Figure 2-9*. As observed in the overt condition, brain areas involved in imagined spectrotemporal decoding were also concentrated around STG, pre and post central gyri.

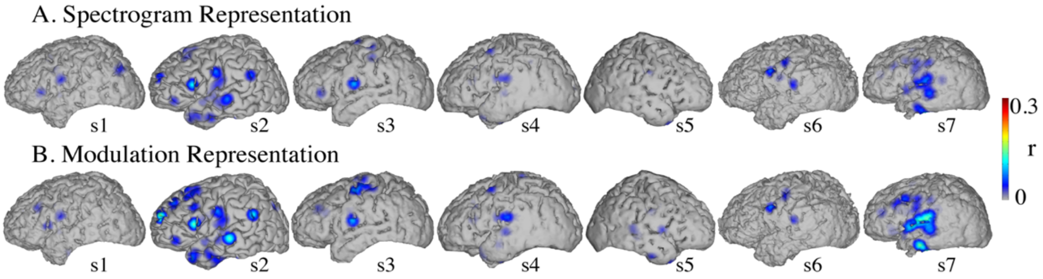


Figure 2-9 Imagined speech predictive power. *Reconstruction accuracy correlation coefficients were computed separately for each individual electrode and for both imagined and baseline control conditions. The plotted correlation values are calculated by subtracting the correlation during baseline control from the imagined condition. The informative area map was thresholded to p<0.05 (Bonferroni correction)* ***(A)*** *Spectrogram-based reconstruction accuracy* ***(B)*** *modulation-based reconstruction accuracy.*

2.5 Discussion

We evaluated a method to reconstruct overt and imagined speech from direct intracranial brain recordings. Our approach was to first build a neural decoding model from self-generated overt speech, and then to evaluate whether this same model could reconstruct speech features in the imagined speech condition at a level of accuracy higher than chance. Our results indicated that auditory features of imagined speech could be decoded from models trained from an overt speech condition, providing evidence of a shared neural substrate for overt and imagined speech. However, comparison of reconstruction accuracy in the two conditions also revealed important differences between overt and imagined speech spectrotemporal representation. The predictive power during overt speech was higher compared to imagined speech and this difference was largest in STG sites consistent with previous findings of a partial overlap of the two neural representations (Geva et al. 2011; Pei et al. 2011; Shuster and Lemieux 2005; Huang, Carr, and Cao 2002). In addition, we compared the quality of the reconstructions by assessing how well they could be identified. The quality of overt speech reconstruction allowed a highly significant identification, while in the imagined speech condition, the identification was only marginally significant. These results provide evidence that continuous features of imagined speech can be extracted and decoded from ECoG signals, providing a basis for development of a brain-based communication method for patients with disabling neurological conditions. In addition, this technique provided a quantitative comparison of the similarity between auditory perception and imagery in terms of neural representations based on acoustic frequency and modulation content.

The relationship between overt and imagined speech reconstruction depended on anatomy. High gamma activity in the superior temporal gyrus, pre- and post-central gyrus provided the highest information to decode both spectrogram and modulation features of overt and imagined speech. However, the predictive power for imagined speech was weaker than for overt speech. This is in accordance with previous research showing that the magnitude of activation was greater in overt than in imagined speech in some perisylvian regions (Palmer et al. 2001; Pei et al. 2011; Partovi et al. 2012) possibly reflecting a lower signal-to-noise ratio (SNR) for HG activity during imagined speech.

While promising, these results highlight the difficulty in applying a model derived from overt speech data to decode imagined speech. This also indicates that the spectrotemporal neural mechanisms of overt and imagined speech are partly different, in agreement with previous literature (Basho et al. 2007; Aleman 2004; Shuster et al. 2005; Pei et al. 2011). Despite these difficulties, it is possible that decoding accuracy may be improved by attention to several factors. First, a major difficulty in this approach is the alignment of imagined speech reconstructions to a reference speech segment. Variability in speaking rate, pronunciation, and speech errors can result in suboptimal alignments that may be improved by better alignment algorithms or by more advanced automatic speech recognition techniques (e.g., Hidden Markov Models). Second, a better scientific understanding of the differences between overt and imagined speech representations may provide insight into how the decoding model can be improved to better model imagined speech neural data. For example, the current study uses a simple model that assumes the auditory representation of imagined speech is equivalent to that of overt speech. If systematic differences in spectrotemporal encoding can be identified during imagined speech, then the spectrotemporal tuning of the decoding model can be biased to reflect these differences in order to optimize the model for imagined speech data. Further investigation of the differences in overt and imagined spectrotemporal neural representation offers a promising avenue for improving imagined speech decoding. Third, the current data sets were obtained with limited training which we predict will have a major impact on reconstruction accuracy. Fourth, improvement in recording electrodes will permit recording of ECoG activity at increasing spatial resolution increasing the information needed for improved speech reconstruction.

Chapter 3 Decoding phoneme sequences during imagined speech

Disclaimer: This chapter is adapted from the following article – with permissions of all co-authors and journals:

Martin S., Bellier L., Lee K., Brunner P., Schalk G., Millán J. d. R., Knight R.T., Pasley B.N. 2014. "Decoding phoneme sequences during imagined speech" in preparation.

My contribution: Conceptualization, formal analysis, methodology, visualization, writing – original draft preparation

3.1 Abstract

Behavioral studies suggest that some forms of phonemic representations occur during imagined speech. However, the brain mechanisms underlying phonemic encoding remain largely unknown. In this study, we investigated the neural correlates of phonemic representations during imagined speech, and evaluated the ability to decode continuous sequences of phonemes from electrocorticographic signals in two patients undergoing neurosurgical procedures for epilepsy. We also compared the results with those obtained during word perception and overt speech production. In order to label intended phonemes more accurately during imagined speech, we designed a karaoke-like task, in which visual words scrolling on the screen were divided into their phonemic representations. Then, we implemented a hidden Markov model (HMM), in order to decode the most likely phoneme sequence given the high gamma (70-150H) neural features and the phonemic transition probabilities. In both participants, phoneme decoding accuracy was significantly higher than chance across eight phonemes in the listening (mean L =27%) and in the overt speech condition (mean O =19%; $p<0.0001$). In the imagined speech condition, the decoding accuracy was only significant in on participant (mean I = 16%; $p<0.05$). Brain areas involved in the classification were located to the middle and superior temporal gyrus, sensorimotor cortex and inferior frontal gyrus – regions typically associated with speech. Using word-specific HMMs, we were able to identify words from the brain activity in the listening (mean=0.62) and overt speech (mean=0.63), but not in the imagined speech (mean=0.50), delineating a number of key challenges.

3.2 Introduction

Early auditory systems decompose complex speech sounds into frequency components. Subsequent steps in the processing stream include the extraction of invariant elements of speech from acoustic features. The superior temporal gyrus plays an important role in transforming these acoustic cues into categorical speech units (Mesgarani et al. 2014). Similarly, the ventral sensorimotor cortex organizes phonetic representations along acoustic features during speech perception (Cheung et al. 2016), and somatotopically based on articulators (lips, tongue, larynx and jaw) during speech production (Bouchard et al. 2013; Cheung et al. 2016). Although much is known about the neural correlates of speech perception and production, it remains unclear if the human brain encodes phonetic features during imagined speech. In this study, we investigated how phonemes, the smallest units of speech, are encoded in the human brain, and evaluated the ability to decode phoneme sequences during imagined speech.

Behavioral studies have provided evidence that phoneme substitution errors occurred between phonemes that shared similar features during both overt and imagined speech (phonemic similarity effect; Corley, Brocklehurst, and Moat 2011). In addition, brain imaging studies have revealed anatomical brain regions involved in silent articulation, such as the sensorimotor cortex, the inferior frontal gyrus, and temporo-parietal brain areas (Pulvermuller et al. 2006). Recently, electrophysiological studies have shown that the neural activity of a listener that perceives a specific phoneme that has been acoustically degraded, replaced or masked by noise is grounded into acoustic representations (Leonard et al. 2016; Holdgraf et al. 2016). This phenomenon, called the phonetic masking effect shows that even in the absence of a given speech sound, the neural patterns correlate with those that would have been elicited by the actual speech sound. These findings suggest that phonemes are represented during imagined speech in the human cortex. From a decoding perspective, several studies have succeeded in classifying individual imagined speech units into different categories, such as covertly articulated vowels (Ikeda et al. 2014), vowels and consonants during covert word production (Pei et al. 2011) and intended phonemes (Brumberg et al. 2011). These studies represent a proof of concept for basic decoding of individual speech units, but failed to incorporate sequential properties during continuous speech.

In this study, we evaluated the ability to decode continuous phoneme sequences from electrocorticographic neural activity, recorded while patients imagined words. We also compared the results with those obtained during word perception and overt production. However, phonemes are embedded in continuous streams of natural speech, and their boundaries are not easily delineated in the physical acoustic signal. This problem becomes even more difficult when investigating imagined speech due to the lack of any behavioral output. In order to label phonemes more accurately, we designed a novel protocol similar to a karaoke-like task, in which visual words scrolling on the screen were divided into their phonemic representations – alternating dark and light colors to indicate phoneme boundaries. Then, we implemented a hidden Markov model (HMM) – a statistical model that allows predicting the most likely phoneme sequence, given the neural activity and a language model. HMMs have been widely used in speech recognition (Rabiner 1993), and more recently applied to decode phoneme sequences from neural patterns during speech perception (Moses et al. 2016) and production (Herff et al. 2015). In the listening and overt speech, classification accuracy was significant in both participants, whereas in the imagined speech, preliminary results were only significant in one participant. The participant that did not reach significance had trouble following the pace of the task, and showed discrepancies between the intended and actual phonemes in the overt speech. Brain areas involved in the decoding process were located to perisylvian brain areas, sensorimotor and inferior frontal areas. In addition to low level phoneme decoding, we were able to identify words in the listening and overt speech, but not in the imagined speech. These findings highlight the difficulty when decoding imagined speech, and delineates a number of key challenges.

3.3 Material and methods

3.3.1 Subjects and data acquisition

Electrocorticographic (ECoG) signals were recorded using subdural electrodes implanted in two patients undergoing neurosurgical procedures for epilepsy. Both patients volunteered and gave their informed consent (approved by the Albany Medical College Institutional Review Board and the

University of California, Irvine and Berkeley Institutional Review Boards and Committees on Human Research) before testing. Electrode grids had center-to-center distance ranging between 4-10 mm. Grid placement and duration of ECoG monitoring were based solely on the requirements of the clinical evaluation (*Figure 3-1*). Localization and co-registration of electrodes was performed using the structural MRI and CT scans. Three-dimensional cortical models of individual subjects were generated using pre-operative structural magnetic resonance (MR) imaging. These MR images were co-registered with the post-operative CT images to identify electrode locations.

Multi-electrode ECoG data were amplified and digitally recorded with sampling rate 3,052. ECoG signals were re-referenced to a common average after removal of electrodes with epileptic artifacts or excessive noise (including broadband electromagnetic noise from hospital equipment or poor contact with the cortical surface). In addition to the ECoG signals, the audio output of the microphone and headphones were recorded along with the multi-electrode ECoG data. This procedure was similar to our previous studies (Martin et al. 2014, 2016, 2017).

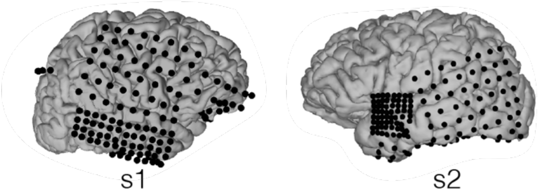


***Figure 3-1 Electrode location**. Grid locations overlaid on each participant's the cortical surface reconstructions.*

3.3.2 Experimental paradigms

We used a word repetition task (overt and imagined) cued with an auditory-visual stimulus presentation. For each trial, words were visually displayed on the screen moving from right to left at the vertical center of the screen (*Figure 3-2 A*). A trial consisted of a sequence of 1) listening to the word (word displayed in blue), 2) one overt speech repetition (word displayed in red) and 3) two imagined speech repetitions (words displayed in orange). In-between words, there was 1sec resting state periods. Words displayed on the screen were divided into their phonemic representations – alternating dark and light colors to indicate phoneme boundaries. The timing of the audio in the listening condition matched the visual cue at the horizontal center of the screen (*Figure 3-2 B; left panel*). Visual cues helped pacing the subject and to repeat the phoneme sequence as close as possible as in the listening condition (*Figure 3-2 B; middle and right panels*). Thus, patients had to repeat (out loud or silently) a given phoneme, when its visual cue reached the horizontal center of the screen, similar to the listening condition. For all three conditions, we recorded the output of the headphones and microphone, together with the intended phoneme state as defined by the visual cue.

The set of phoneme stimuli was carefully chosen in order to maximize chances of discriminating brain patterns associated with each type. To achieve this, we selected phonemes that varied in place and manner of articulation. Stop consonants, in which the vocal tract is blocked and all airflow ceases, were not considered. Instead, we took consonants that could be sustained for the time indicated by the visual cue, such as nasal (n), fricative (sh, s, and v) and approximant (l) consonants. Vowels were selected in order to have different tongue positions such as ah (open back), eh (mid central), ih (close front). The final set consisted of eight phonemes and one additional silence phoneme (sp), coming from the resting state intervals.

Nine words were analyzed in this task, which were selected based on the phoneme set (shovel, novel, seven, shin, sleeve, slosh, olive, illness, envy). Trials were randomly ordered and repeated once per run. Words had between 3 and 5 phonemes and lasted about 4sec. Each run lasted about 4 min, and was repeated 3-5 times depending on the health condition of the participants. Auditory stimuli were recorded from a native American speaker, and then stretched in time to adjust the timing of the task. This means that the temporal structure of the words was maintained, and phonemes could have different time length across different words. The precise task design and timing is summarized in *Figure 3-2*. The microphone recording was used to verify that subjects were not producing audible speech during imagery, as well as monitoring the behavior (speech delays and word length) during overt speech. The intended phoneme label was recorded in synchrony with the ECoG neural data – allowing marking the neural data and associate the neural patterns and phonetic content at each given time.

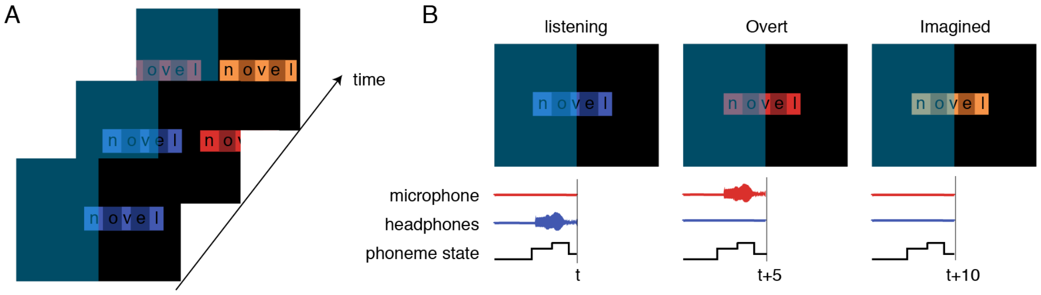


Figure 3-2 Task design. ***(A)*** *Words were displayed on the screen alternating dark and light colors to indicate phoneme sequences, in three conditions: 1) listening to the word, 2) one overt speech repetition and 3) two imagined speech repetitions.* ***(B)*** *In the listening condition (left panel), the audio matched the visual cue at the horizontal center of the screen. In the overt speech condition (middle panel), patients had to repeat out loud a given phoneme, when its visual cue reached the horizontal center of the screen, as close as possible as in the listening condition. I the imagined speech condition patients had to repeat silently a given phoneme, when its visual cue reached the horizontal center of the screen. In this example, at time t, the participant was hearing the phoneme v. About 5sec later, he had to repeat out loud phoneme v, whereas 10sec later, he had to repeat silently phoneme v.*

3.3.3 Feature extraction

High gamma activity was computed using eight bandpass filters (Butterworth filter of order 4, logarithmically increasing center frequencies (70–150 Hz)), and extracted the envelope using the Hilbert transform. The power was then calculated by averaging the signal across these eight bands. Subsequently, the signal was down-sampled to 100 Hz and z-scored (Bouchard et al. 2013). The microphone, headphones and intended phoneme trigger channel were also downsampled at 100 Hz.

3.3.4 Feature selection

Features were selected on the training set. The first step in the process consisted in selecting electrodes with significant high gamma activity during phonemes perception/production. For this, we extracted phonemes based on the visual cue (between 117-290 phoneme trials); this corresponded to the actual phoneme auditory stimuli in the perception condition and the intended phoneme production in the overt and imagined conditions. In order to compare phonemes with different durations, we normalized the time scale. Then, we took the mean across trials and divided by the standard deviation. Finally, we averaged across time in order to have a quantitative measure of activation for each electrode and phoneme class. Examples of phoneme activation time courses are displayed in *Figure 3-4,* together with the average activation map *(Figure 3-5).* For each cross-validation fold, electrodes that were above the 95th percentile of the distribution of activation metrics were selected, and concatenated into one feature vector $\boldsymbol{x_t}$, where t represents time samples.

In order to integrate temporal dynamics into the classification framework, we included time lags up to 500ms into the feature map, $\boldsymbol{v_t} = [\boldsymbol{x_t}, \boldsymbol{x_{t+1}}, \dots, \boldsymbol{x_{t+500}}]$. Finally, to reduce the number of features, we computed Fisher's score (Fisher 1936), a discriminability index based on the ratio of between class scatter to within class scatter. For each feature, the between-class scatter is defined as:

$$SB = \frac{1}{J}\sum_{j=1}^{J}(\mu_j - \mu)(\mu_j - \mu)^T$$

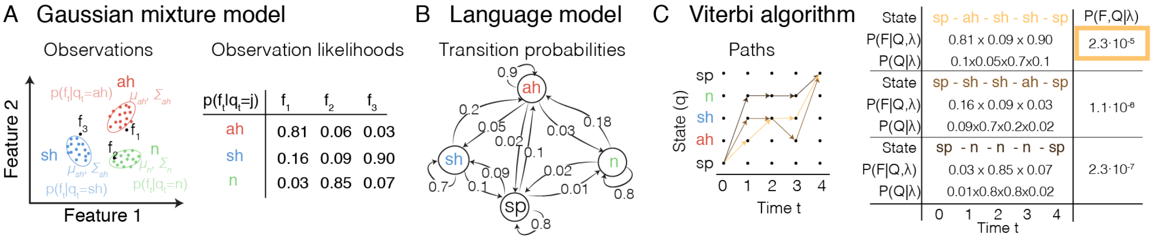
where μ_j is the mean of class j, and μ is the mean of the class means. The between-class scatter is defined as:

$$SW = \sum_{j=1}^{J}\Sigma_j$$

Σ_j is the variance of class j. The reduced neural features $\boldsymbol{f_t}$ are a subset of the original data $\boldsymbol{v_t}$, defined using inner loop cross-validation. For this, we used forward sequential feature selection until there is no improvement in prediction on the validation set. The feature set that led to the best validation accuracy averaged across inner folds was selected. The final input data $\boldsymbol{F} \in \mathbb{R}^{PxT}$, where P is the number of high gamma features selected and T is the number of time samples. The number of selected features depends on the inner loop cross-validation and changes for each fold.

3.3.5 Classification

In continuous hidden Markov models, the states are not directly observed (hidden), but observations do depend on them. In our case, we tried to recover the sequence of hidden states (phoneme labels), from the observation states (continuous neural feature space). In this framework, the most likely sequence of phonemes was estimated by incorporating two probabilistic models: a Gaussian mixture model (*Figure 3-3 A*) and a language model (*Figure 3-3 B*; Rabiner 1993). The Gaussian mixture model evaluated the likelihood of each phoneme at any time point, given the neural features. The language model defined the phoneme transition probabilities. Finally, in order to find the most likely phoneme sequence given the combination of neural observations and both models, we used a Viterbi decoding algorithm (*Figure 3-3 C*).



p(f$_t$|q$_t$=j)	f$_1$	f$_2$	f$_3$
ah	0.81	0.06	0.03
sh	0.16	0.09	0.90
n	0.03	0.85	0.07

State	sp - ah - sh - sh - sp	P(F,Q|λ)
P(F|Q,λ)	0.81 x 0.09 x 0.90	2.3·10^{-5}
P(Q|λ)	0.1x0.05x0.7x0.1	
State	sp - sh - sh - ah - sp	
P(F|Q,λ)	0.16 x 0.09 x 0.03	1.1·10^{-6}
P(Q|λ)	0.09x0.7x0.2x0.02	
State	sp - n - n - n - sp	
P(F|Q,λ)	0.03 x 0.85 x 0.07	2.3·10^{-7}
P(Q|λ)	0.01x0.8x0.8x0.02	
	0 1 2 3 4 Time t	

Figure 3-3 Elements of the decoding framework. (A) *Gaussian mixture models implementing the likelihood of observing the neural activity given a phoneme type. Feature 1 and Feature 2 represent two dimensions of the neural feature. Each phoneme type is modeled by a multivariate Gaussian model with mean* $\boldsymbol{\mu_j}$ *and covariance* $\boldsymbol{\Sigma_j}$*;* $\boldsymbol{f_1}$*,* $\boldsymbol{f_2}$ *and* $\boldsymbol{f_3}$ *represent examples of three neural feature vectors at time t=1, t=2 and t=3,*

respectively. The right panel depicts the likelihoods of each neural feature vector to be in a given phoneme state ("ah", "sh" or "n"). ***(B)*** *Language model defining the probabilities of going from one phoneme state to another. Numbers described here do not represent true English probabilities, but are just for illustrative purposes.* ***(C)*** *Viterbi algorithm that finds recursively which is the most likely sequence of state given the observations and the HMM. Three examples of paths are showed, and the one with the largest joint probability $P(F, Q|\lambda)$ is chosen.*

In addition to low level phoneme decoding, we also evaluated the ability to identify words from the decoded phoneme units. For this, we built one HMM for each type of word (N=9), and computed the probability that the sequence was produced by each model. The model with the best fit to the sequence of phonemes was selected (see section *3.4.4. Word identification*).

Finally, given the difficulty in decoding phoneme sequences during imagined speech, we also evaluated the ability to distinguish between speech and silence states. For this, we used the same feature selection and Gaussian mixture models as in the phoneme-based decoding, but modeled only two states: speech on (all phonemes) and speech off (silence states) (see section *3.4.2. Speech detection*).

Gaussian mixture model

The Gaussian mixture model (GMM) defined the most likely phoneme (hidden state) type given the neural feature (continuous observation space). For this, we used multivariate Gaussian densities to model each class using the training set. We modeled each phoneme j with a multivariate normal distribution model with mean vector $\boldsymbol{\mu_j}$ and covariance $\boldsymbol{\Sigma_j}$, computed from the neural feature training vectors associated with the class j. As such, the likelihood of the neural activity feature vector $\boldsymbol{f_t}$ given that the phoneme state is $q_t = j$ followed a normal distribution:

$$P(f_t|q_t = j) \sim \mathcal{N}(\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$$

This likelihood measure was then converted to a posterior probability that a neural feature vector was produced by the model $\mathcal{N}(\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$ using Bayes rule, such as:

$$P(q_t = j|f_t) \propto P(f_t|q_t = j) \cdot P(q_t = j)$$

where $P(q_t = j)$ is the prior probabilities of the being in state j. These priors were calculated from the relative proportion of each class in the training set, and changed across cross-validation folds. We normalized the posterior probability distributions as follows:

$$b_j(f_t) = P(q_t = j|f_t) = \frac{P(f_t|q_t = j) \cdot P(q_t = j)}{\sum_k P(f_t|q_t = k) \cdot P(q_t = k)}$$

$\mathrm{b_j}$ represents the emission probabilities of HMM, and $k \in [1, \ldots, K], K = 8$ phoneme classes. The feature vector f_t was assigned to the phoneme that had the highest posterior probability:

$$\underset{j}{\operatorname{argmax}}\big(b_j(f_t)\big)$$

Language model

The language model provides the a priori phoneme transition probabilities (Rabiner 1993). Typically, these probabilities are estimated on large corpora containing phoneme sequences. In this study, due to the small and specific dictionary size, we calculated the transition probabilities on the training label sequence. Given the limited amount of available data, co-articulations were not modeled (context-independent), and we represented each phoneme by a single HMM state, i.e. eight phoneme states. The probabilities of going from one phoneme state to another were defined as follows:

$$a_{ij} = P(q_{t+1} = j|q_t = i)$$

where q_t is the actual phoneme state at time t that can be any of the eight phoneme type, $a_{ij} \geq 0$ and $\sum_j a_{ij} = 1, \forall i$. The initial state probabilities represent the probability to start the sequence ($t = 1$) with a given state:

$$\pi_i = P(q_1 = i)$$

In this study, we used a first order Markov model, in which the probability of being at a given state at time t depends only on the state at time $t - 1$.

Viterbi algorithm

The problem to solve in this task is, given the sequence of neural feature vectors $\boldsymbol{F} = (\boldsymbol{f_1}, \boldsymbol{f_2}, \dots, \boldsymbol{f_T})$ and the model $\lambda = \{\pi_i, a_{ij}, b_j\}$, what is the corresponding hidden state sequence $Q = (q_1, q_2, \dots, q_T)$ that best explains the observations. The probability of observing the sequence F given the hidden state sequence Q and the model λ is defined as:

$$P(F|Q,\lambda) = \prod_t^T P(q_t = j|f_t) = \prod_t^T b_j(\boldsymbol{f_t})$$

Whereas the probability of having the sequence Q given the model λ.

$$P(Q|\lambda) = \prod_t^T P(q_t = j|q_{t-1} = i) = \pi_{q_0} \prod_t^T a_{q_{t-1}q_t}$$

Finally, the probability of observing the sequence F, given the sequence Q and the model λ is given by the joint probability:

$$P(F, Q|\lambda) = P(F|Q,\lambda) \cdot P(Q|\lambda) = \prod_t^T b_j(\boldsymbol{f_t}) \cdot a_{q_0} \prod_t^T a_{q_{t-1}q_t}$$

We used the Viterbi algorithm for searching the optimal state transition sequence (Viterbi path), that could have generated a given output sequence. The Viterbi algorithm is a dynamic programming algorithm that was conceived by Andrew Viterbi in 1967 as an error-correction scheme for noisy digital communication links (Forney 1973). This search algorithm finds the least cost effective path in a grid, by first evaluating the cost to reach every state in the grid, and then tracing back the path that corresponded to the overall most likely sequence. The phoneme sequence with the highest joint probability was selected:

$$Q^* = \underset{Q}{\operatorname{argmax}} P(F, Q|\lambda)$$

Word identification

We built a specific model $\lambda_w = \{a_i, a_{ij}, b_j\}$ per word ($W = 9$), and determined the overall likelihood of a neural observation sequence $\boldsymbol{F} = (\boldsymbol{f_1}, \boldsymbol{f_2}, \dots, \boldsymbol{f_T})$ being generated by each model. The predicted word is the one with the highest likelihood among all the words in the candidate set.

3.3.6 Evaluation

Given data limitations, we performed leave-one-word-out cross-validation. For each condition, the feature selection and model fitting were performed on the training set only, and evaluated on the independent testing set. To define the accuracy of the system, several metrics were used. First, we report the overall classification accuracy as the average number of correctly classified phoneme over the sequence. In addition, in order to identify if the system successfully predicted each phoneme type, or if the classification was biased towards one class, we computed the confusion matrix between the actual and predicted phonemes for each cross-validation fold. We normalized the confusion matrices by the number of samples in each class, in order to have the rate of classification accuracy for each phoneme type. The phoneme confusion accuracy was the mean of the diagonal elements, averaged across folds. This metric weighted the classification accuracy for each phoneme equally, accounting for the unbalanced number of samples in each class. Finally, we computed the Area Under the Curve (AUC), in order to determine if the true positive rate (diagonal elements of the confusion matrix) was higher than the false positive rate (off-diagonal elements). An AUC coefficient of 1 indicates that the classifier predicts well all the samples, whereas an AUC of 0.5 corresponds to random predictions

We also evaluated the ability to identify words from the decoded phoneme sequences, by computing the identification rank. To this end, the likelihood was computed for each word in the candidate set, using the specific HMMs and the neural observation sequences. Then, these likelihood scores were sorted, and the word identification rank was defined as the percentile of the correct word. An identification rank of 1 indicates that actual word has the highest likelihood among all the words in the candidate set, whereas an identification rank of 0 indicates that the actual word has the lowest likelihood. At chance level, the mean identification rank is 0.5.

3.3.7 Statistics

Theoretical chance level in an eight-class problem is 12.5%. However, in order to compare how well the language model performed without any neural information, we computed the true random level and evaluated statistical significance using randomization tests. For this, we shifted circularly 200 times the training labels by a random integer, thus maintaining the phonemic transition probabilities of the actual model. Then, we performed the exact same feature selection and classification steps as in the actual classification process. The averaged accuracy across cross-validated testing set yielded one value in the null distribution. We tested if the mean of the surrogate null distribution was significantly different from the theoretical chancel level using Student's t-test. The mean distribution was not significantly different from the theoretical chance level ($p>0.5$; two-sided one-sample t-test), and we therefore kept the theoretical level to compute our statistics. We computed the p-value that the performances across leave-one-word-out folds was higher than chance levels using non-parametric Wilcoxon signed rank test, and applied the FDR correction for multiple comparison.

We performed similar randomization tests to define chance levels in the speech detection task. In this two-class approach, the mean distribution was not significantly different from the theoretical chance level of 50% ($p>0.5$; two-sided one-sample t-test), and we therefore kept the theoretical level to compute our statistics.

3.4 Results

3.4.1 High gamma neural activity

We analyzed high gamma temporal dynamics at different electrode locations, and compared brain patterns across conditions (listening, overt, and imagined speech) for the different phoneme types. For this, we extracted trials for each phoneme class, normalized trials in the time domain in order to deal with trials of different time length, and standardized time samples across trials. Normalized activation time courses are depicted in *Figure 3-4* for different electrodes. Neural activation patterns revealed distributed spatiotemporal brain responses across electrodes, phonemes and experimental conditions, thus highlighting the complexity of the dynamical system associated with speech processing. As such, it is difficult to precisely define selectivity for phonetic features from these results.

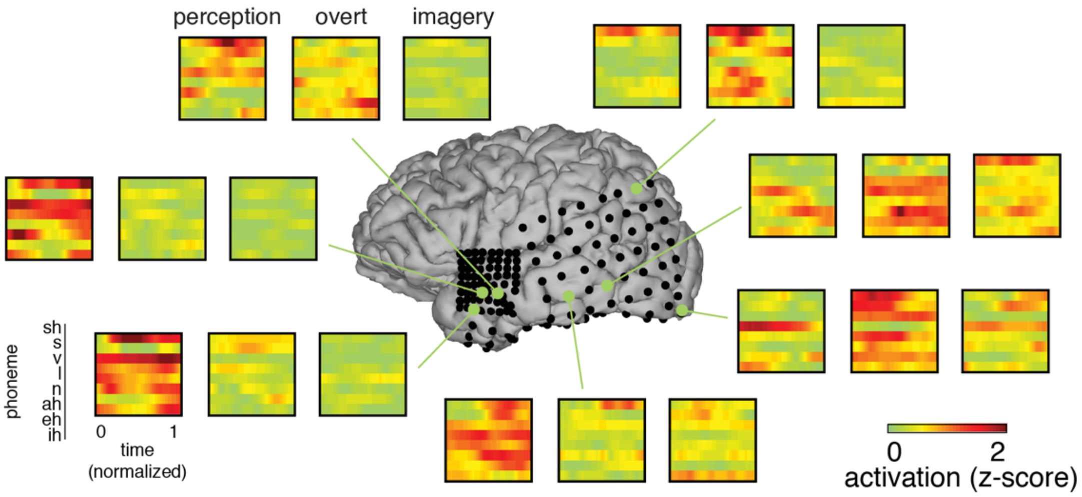


***Figure 3-4 High gamma neural activation. (A)** Examples of high gamma standardized neural activity across trials for each phoneme (arpabet) and condition.*

In order to have a more representative overview of the activation map, we plotted the average brain activation across time and phonemes on the surface rendering of the patients' cortex (*Figure 3-5*). Results showed activations in perisylvian areas during all three conditions. The sensorimotor cortex was active in the overt speech conditions, whereas the inferior frontal cortex in both listening and overt speech conditions. These are brain areas typically involved in speech processing. To assess similarities across conditions, we computed the correlation coefficient between the activation patterns of the different conditions. Results showed that the neural response in the imagined condition correlated to various degrees with listening and overt speech conditions across subjects. This may be explained by the various strategies employed to produce inner speech (e.g. auditory versus kinesthetic imagery), although we tried to control this aspect with the instructions given to the patients. These results revealed that the brain representations underlying the different conditions are partially overlapping and partially dissociable.

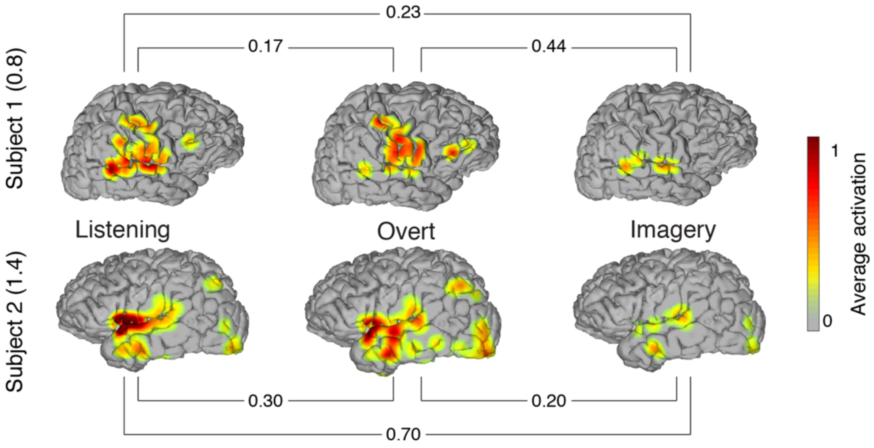


Figure 3-5 Neural activation map. *Brain activation patterns averaged across time and phonemes, and plotted on the surface rendering of the patients (Map thresholded at p < 0.05). Each patient is scaled to the maximum value across all conditions (indicated by the number into parenthesis). We computed the correlation between the activation patterns of the various conditions.*

3.4.2 Speech detection

We evaluated the ability to distinguish between speech and silence states in all three conditions (listening, overt and imagined speech). In order to take into account unbalanced data, we computed the average accuracy, as the mean between true positive rate and true negative rate. For both participants, results showed that the classification accuracy was significantly higher than chance levels in the listening (mean L = 75%) and overt speech (mean O = 72%; $p<10^{-4}$; one-sided one-sample Wilcoxon signed rank test; *Figure 3-6* A). In the imagined speech condition, the classification accuracy was significant in subject 2 (mean I = 61%; p<0.01), but not significant in subject 1 (mean I = 55%; p>0.1; one-sided one-sample Wilcoxon signed rank test; FDR correction). The bias of the classifier across subject and condition was less than 3%. These results suggest that high gamma activity detects imagined speech states, and could be used as a biomarker defining an active speech window. However, it also emphasizes the difficulty when tackling internal processes, such as imagined speech.

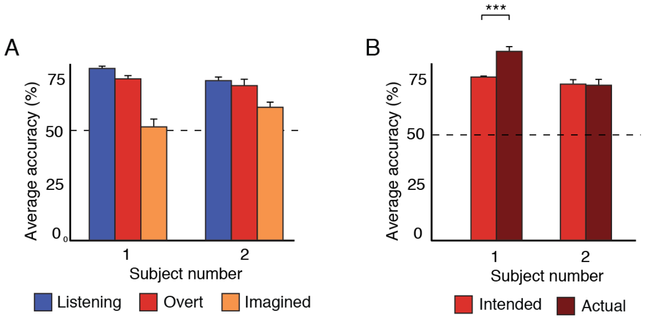


Figure 3-6 Speech detection. *(A) Class average accuracy for all condition and subjects. Chance level is 50%* ***(B)*** *Comparison of the classification accuracy in the overt speech condition – when models were fit with the intended speech labels (visual cue) or with the actual speech labels (labeled manually from the recorded speech). Error bars denote standard error of the mean.*

One major issue in the overt speech condition was the discrepancies between actual speech and intended speech (defined by the visual cue). These inconsistencies represented a challenge for building the decoding model, as the phoneme labels were not corresponding to the correct neural features. In order to quantitatively measure temporal irregularities across trials, we manually labeled the recorded overt speech, and compared it to the intended speech. The average speech onset delay with respect to the visual cues was -0.47s±0.57 and 0.09s±0.15 for subject 1 and subject 2, respectively. Subject 1 was particularly bad at keeping up the pace and following the visual cues, and only 42% of the labels matched between actual and intended speech (89% for subject 2). This may account for the poor imagined decoding in subject 1. Similar behavior were to be expected in the imagined speech condition, as both overt and imagined speech have been shown to be subject to similar speech production temporal variations (Hubbard 2010). In the listening condition, this was not a problem given that the auditory stimuli were time-locked across repetitions, and thus intended phonemes corresponded to actual phonemes.

In order to measure the impact of speech production irregularities, we evaluated possible improvements in the overt speech condition, when the classifier was built on the actual speech rather than on the intended speech. Performances of subject 1 – who was irregular across repetitions – significantly improved with the actual speech (mean O_{actual} = 84%) over the intended speech (mean $O_{intended}$ = 71% ; $p<10^{-4}$; unpaired two-sample Wilcoxon ranksum test; FDR correction; *Figure 3-6 B*). Conversely, the classification accuracy of subject 2 – who was consistent across trials – did not significantly improve (mean O_{actual} = 72%; mean $O_{intended}$ = 73%; $p>0.5$; unpaired two-sample Wilcoxon ranksum test; FDR correction). These results shows that the classification accuracy was degraded proportionally to the level of speech irregularity, which emphasize the importance of having the correct labeling in the current modeling framework.

3.4.3 Phoneme-based decoding

Using continuous density hidden Markov models, we evaluated the ability to decode phoneme sequences from the neural features in three conditions (listening, overt and imagined speech). For each participant, we quantified performances using the overall classification accuracy and the class average accuracy. Given that the silence phoneme may have inflated the classification accuracy, we did not model it in this section. Results showed that the overall classification accuracy was significantly higher than chance level in the listening (mean L = 27%) and in the overt speech (mean O = 19%; $p<0.001$; one-sided one-sample Wilcoxon signed rank test; FDR correction; *Figure 3-7 A*). While the overall classification was significant in the imagined speech for subject 2 (mean I = 16%; $p<0.01$), it did not reach the significance level in subject 1 (mean I = 12%; $p>0.5$). In both participants, the listening condition and overt speech were significantly better than the imagined speech condition ($p<0.01$; unpaired two-sample Wilcoxon ranksum test; FDR correction).

In order to verify that individual phonemes were classified above chance, and that the classifier was not biased towards one class, we computed the confusion matrices. Results are displayed for subject 2 in *Figure 3-7 B* for all three conditions. In the listening and overt speech conditions, the diagonal of the matrix stood out from the rest of the elements, suggesting that phonemes were accurately decoded from high gamma neural features. However, at the individual level, not all phonemes were better than chances; seven and six out of eight phonemes were significant for the listening and overt speech, respectively ($p<0.05$; one-sided one-sample Wilcoxon signed rank test). Conversely, at the group level, all eight phonemes were significantly above chance level for the listening for both participants ($p<0.01$; one-sided one-sample Wilcoxon signed rank test). In the imagined speech condition, phonemes were more likely to be misclassified as reflected in the confusion matrix; only four out of eight phonemes were significantly above chance (p<0.05; one-sided one-sample Wilcoxon signed rank test). The classification accuracy in the imagined speech was significantly better than chance across all phonemes in this participant (p<0.05; one-sided one-sample Wilcoxon signed rank test; FDR correction). AUC coefficients were higher than chance level, for both participants in the listening and overt speech condition (*Figure 3-7 C*; p<0.001). In the imagined speech condition, results were only significant in subject 1 (p<0.05; one-sided one-sample Wilcoxon signed rank test; FDR correction).

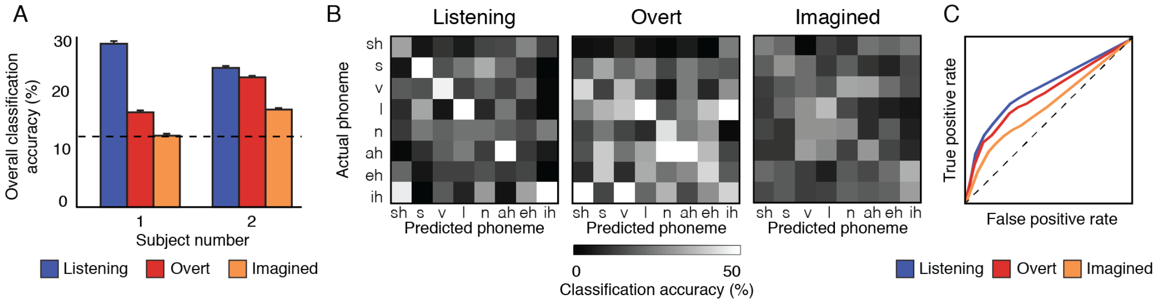


***Figure 3-7 Phoneme classification accuracy. (A)** Overall classification accuracy across leave-one-word-out folds for all three conditions (listening, overt and imagined speech). Error bars represents the standard error of the mean. **(B)** Confusion matrix averaged across folds for the listening, overt and imagined speech, respectively, for S1. **(C)** Area under the curve computed from the confusion matrices for S1, determined if the diagonal elements (correctly classified) were higher than the off-diagonal elements (wrongly classified).*

We evaluated the effect of the language model in the classification framework. For this, we compared the overall classification accuracy of the hidden Markov models (HMMs), which incorporated information about the transition probabilities, with that of the Gaussian mixture model alone (GMMs). By comparing both results, we measured the impact of the language model and Viterbi decoding on the performance of the system. At the individual level, the classification accuracy using the language model was not significantly better than when using only the Gaussian mixture model (p>0.05; paired two-sample Wilcoxon signed rank test; FDR). However, across conditions and subjects, the HMM-based classification was better than the GMM-based classification (p<0.05; paired two-sample Wilcoxon signed rank test)

We evaluated the impact of misalignment between the intended speech and the actual speech, in the overt speech. The average phoneme onset delays with respect to the visual cues were -0.49s±1.01 and 0.03s±0.27 for subject 1 and subject 2, respectively. We quantified possible improvements in the overt speech condition, when the decoding models were built on the actual produced phoneme rather than the intended phonemes. The decoding procedure was similar than in the previous analysis. Similar to what was observed in the speech detection, the improvement of subject 1 was significant with the actual phonemes (mean O_{actual} = 24%) compared to with the intended phonemes (mean $O_{intended}$ = 17%; p<0.05; unpaired two-sample Wilcoxon ranksum test; FDR correction), but not in subject 2 (mean O_{actual} = 22% and mean $O_{intended}$ = 20%; p>0.05; unpaired two-sample Wilcoxon ranksum test; FDR correction).

3.4.4 Word based identification

Finally, we also evaluated the ability to identify words from the high gamma brain activity. Examples of an actual phoneme sequence is shown together with the predictions from the Gaussian mixture models (GMMs) and hidden Markov models (HMMs) for all three conditions (*Figure 3-8 A*). Results show that the predictions from the GMMs are bumpy, sometimes changing phoneme states at every time frame. Such phoneme transitions are unlikely in real, naturalistic speech production. The language model corrects for unlikely transitions, and allows the analysis to remain in a given phoneme state for a longer period of time. This behavior resembles more accurately actual speech properties. For the identification, we computed the likelihood that the decoded phoneme sequence belonged to each word in the candidate set, using word-specific HMMs and the neural observation sequences. These likelihood scores were sorted, and word identification was quantified as the percentile rank of the correct word. Results showed that the identification rank was significant for both subjects in the listening (mean L=0.62) and overt speech (mean O=0.63) conditions (p<0.05; one-sided one-sample Wilcoxon signed rank test; FDR correction *Figure 3-8 B*). Conversely, the identification rank did not reach the significance level (mean I=0.49; p>0.05; one-sided one-sample Wilcoxon signed rank test; FDR correction). This suggests that although the phoneme classification was significant, it remain insufficient for word identification in the imagined speech condition.

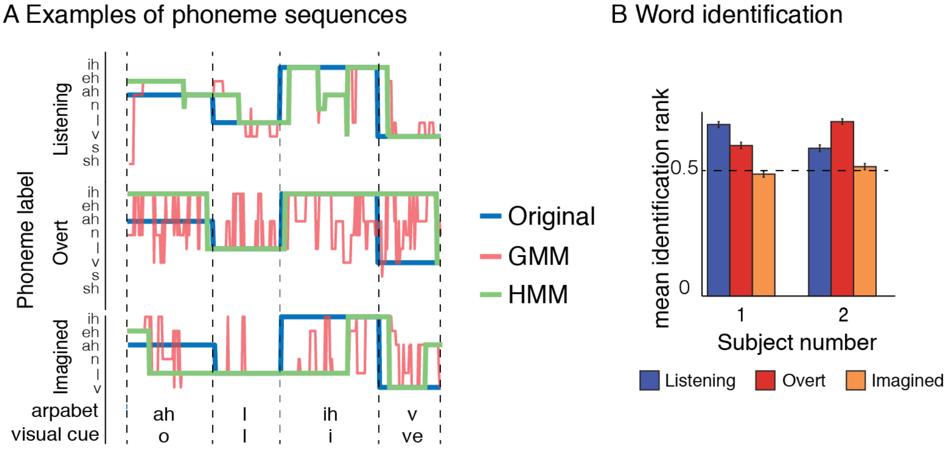


Figure 3-8Word identification. ***(A)*** *Examples of actual phoneme sequences for the word "olive", together with the sequence decoded from the Gaussian mixture model (GMM) and with the hidden Markov model (HMM) for all three conditions.* ***(C)*** *Mean word identification rank.*

3.5 Discussion

In this study, we evaluated the ability to decoding phoneme sequences during imagined speech. By continuously cuing and preparing the participant, we designed a karaoke-like task that sought to reduce the variability in speech production (overt and imagined), and improve the labeling of imagined speech units. We decoded phoneme sequences from high gamma neural activity, using continuous density hidden Markov models – a technique widely used in speech recognition that allows incorporating a language model. We also compared the results with those obtained during word perception and overt production. Results showed that the overall classification accuracy was significant for the listening and overt speech in both participants. In the imagined speech condition, the phoneme-based classification accuracy was only significant in one participant. Although, the language model tended to improve the decoding accuracy compared to phoneme sequences, the classification accuracy was still significant when using solely the Gaussian models. This suggested that discriminant information was embedded in the neural features. From the decoded sequence, we were able to identify individual words in the listening and overt speech condition, but not in the imagery condition. These preliminary results represent a first step towards continuous decoding of phonemes from imagined speech. However, more participants are required to draw any firm conclusions. This preliminary phoneme work emphasizes a number of challenges that were encountered.

For instance, we designed a karaoke-like task, in which words were visually displayed on the screen moving from right to left, and were divided into their phonemic representations. This task design sought to help pacing the patient to be more consistent, and to monitor more precisely the production of imagined speech. However, given the difficulty of one participant (subject 1) to follow up with the pace in the overt speech condition, the task design was not optimal. As a consequence, we observed discrepancies between intended phonemes (from the visual cue) and actual phonemes (manually labeled from the recorded speech), and overall performances were degraded. This was particularly noticeable for subject 1, where the classification accuracy improved, when the model was fit with the actual speech labels rather than with the intended speech labels. Similar temporal irregularities have been observed during imagined speech (Hubbard 2010), which might have also precluded higher performances.

In addition to a novel task design, we implemented a hidden Markov model – a technique that has demonstrated its potential in speech recognition (Rabiner 1993), and more recently in neural-based speech recognition (Moses et al. 2016; Herff et al. 2015). Here, we replicated the ability to decode phoneme sequences during speech perception and production, and extended the approach to imagined speech. However, the classification accuracy was only slightly improved when incorporating the language model compared to when using solely the Gaussian models. This suggests that in this particular case – where only nine words composed of eight different phonemes were decoded – the language model was not relevant.

We found that some phonemes were classified at higher rates than others, and this was directly linked to the amount of data available and possibly other linguistic features. Due to time constraints in the epileptic monitoring unit, we only investigated a small subset of all English phonemes – eight phonemes out of 44. This provided an insight into the phonemic discriminability, but prevented investigating what precise phonetic features were encoded, such as place of articulation (e.g. bilabial, dental) or manner articulation (e.g. nasal, plosive, fricative). Despite the few number of phonemes investigated here, the classification accuracy remained low.

A common hurdle in the field of speech recognition – also faced with neural pattern recognition – lies in the type of speech unit (phoneme, word, sentence) decoded, and the size of the dictionary needed to represent natural, conversational speech. Here, we decoded phonemes, the smallest units of speech. Although, the classification accuracy remained low – for numerous reasons cited above – decoding phonemes has great potential, as it allows building words by stringing units together sequentially. This was demonstrated in the listening and overt speech condition, where words were successfully identified. The advantage of this approach is that it captures the temporal structure and relationship, rather than using a frame-by-frame prediction model. Currently, the small dictionary size (nine words) represents a limitation, and the feasibility for real-time, natural speech decoding awaits further research.

Another impediment to accurate classification is that co-articulations were not modeled in the current study, whereas in natural speech phonemes are influenced by preceding and following sounds. These co-articulations have been shown to be captured in the high gamma neural activity (Bouchard et al. 2013), and could be potentially modeled using more complex HMM structures. Commonly, when the signal is not uniform along its length, each phoneme is modeled with three states, representing the initial part, middle, final part, respectively. While the use of triphones solves the problem of context dependency, it adds complexity to the model. Given the limited amount of data available with ECoG, this could lead to overfitting.

Participants were cued by words scrolling on the screen. As such, this task design did not allow the participant to communicate freely, in terms of what and when to convey a speech instance. Further improvements have to be done in order to decode imagined speech in an asynchronous manner. In this study, results suggested that speech was discriminated from silence states, suggesting that high gamma neural activity could provide a reliable biomarker of speech production during imagined speech.

Chapter 4 Word classification during imagined speech

Disclaimer: Most of this chapter is adapted from the following article – with permissions of all co-authors and journals:

Martin S., Brunner P., Iturrate I., Millán J.d.R., Schalk G., Knight R.T., Pasley B.N. 2016. "Word Pair Classification during Imagined Speech Using Direct Brain Recordings." Scientific Reports. doi:10.1038/srep25803.

My contribution: Conceptualization, formal analysis, methodology, visualization, writing – original draft preparation.

4.1 Abstract

In this study, we evaluated the ability to identify individual words in a binary word classification task during imagined speech, using high gamma (70-150Hz) features in the time domain. For this, we used an imagined word repetition task cued with a word perception stimulus, and followed by an overt word repetition, and compared the results across the three conditions. We used support-vector machines, and introduced a non-linear time-realignment in the classification framework – in order to deal with speech temporal irregularities. As expected, high classification accuracy was obtained in the listening (mean=89%) and overt speech condition (mean=86%), where speech stimuli were directly observed. In the imagined speech condition, where speech is generated internally by the patient, results show for the first time that individual words in single trials were classified with statistically significant accuracy. Classification accuracy reached 88% in a two-class classification framework, and average classification accuracy across fifteen word-pairs was significant across five subjects (mean=58%). The majority of electrodes carrying discriminative information were located in the superior temporal gyrus, inferior frontal gyrus and sensorimotor cortex, regions commonly associated with speech processing. These data represent a proof of concept study for basic decoding of speech imagery, and delineate a number of key challenges to usage of speech imagery neural representations for clinical applications.

4.2 Introduction

People with speech production impairments would benefit from a system that can infer intended speech directly from brain signals. Here, we used direct cortical recording to examine if individual words could be selected during imagined speech within a binary classification framework. However, despite intense investigation, the neural mechanisms underlying imagined speech remain poorly defined in part due to the lack of clear timing of inner speech, the subjective nature and inter-individual differences in how subject imagine speech. Functional magnetic resonance imaging studies have shown that imagined speech activates Wernicke's area (Yetkin et al. 1995; McGuire et al. 1996; Palmer et al. 2001; Shergill et al. 2001; Aleman 2004; Aziz-Zadeh et al. 2005; Geva, Correia, and Warburton 2011) and Broca's area (Hinke et al. 1993; Huang, Carr, and Cao 2002) – two essential language areas involved in speech comprehension and production, respectively (see Price 2012; Perrone-Bertolotti et al. 2014 for reviews). Although traditional brain imaging techniques have identified anatomical regions associated with imagined speech, these methods lack the temporal resolution to investigate the rapid temporal neural dynamics during imagined speech (Towle et al. 2008). In contrast, electrocorticography is a direct neural recording method that allows monitoring brain activity with high spatial, temporal, and spectral resolution (Ritaccio et al. 2014).

In this study, we took advantage of the high resolution offered by ECoG classify individual during imagined speech, using HG features in the time domain. However, speech production (both overt and imagined) is subject to temporal variations (speech onset delays and local stretching/compression) across repetitions of the same utterance (Rabiner 1993; Vaseghi 2007). As a result, a classifier that assumes fixed time features may not recognize two trials as belonging to the same class if the neural patterns were not temporally aligned. To overcome that limitation, we proposed a new classification framework that accounted for temporal variations during speech production (overt and imagined) by introducing time realignment in the feature map generation. In particular, we used support-vector machines (Hastie 2009) to classify individual words in a word pair, and introduced a non-linear time alignment into the kernel to deal with internal speech production variability. We used an imagined word repetition task cued with a word perception stimulus, and followed by an overt word repetition, and compared the results across the three conditions (listening, overt and imagined speech). As expected, high classification accuracy was obtained in the listening and overt speech condition where speech stimuli were directly observed. In the imagined speech condition, where speech is generated internally by the patient, results show for the first time that individual words in single trials were classified with statistically significant accuracy. The majority of electrodes carrying discriminative information were located in the superior temporal gyrus, inferior frontal gyrus and sensorimotor cortex – regions commonly associated with speech processing. Notably, the most robust decoding effects were observed in the temporal lobe electrodes.

4.3 Material and methods

4.3.1 Subjects and data acquisition

Electrocorticographic (ECoG) recordings were obtained using subdural electrode arrays implanted in 5 patients undergoing neurosurgical procedures for epilepsy. All patients volunteered and gave their informed consent (experimental protocol was approved by the Albany Medical College Institutional Review Board and methods were carried out in accordance with the approved guidelines and regulations) before testing. The implanted electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) consisted of platinum–iridium electrodes (4 mm in diameter, 2.3 mm exposed) that were embedded in silicon and spaced at an inter-electrode distance of 4-10 mm. Grid placement and duration of ECoG monitoring were based solely on the requirements of the clinical evaluation (*Figure 4-1*).

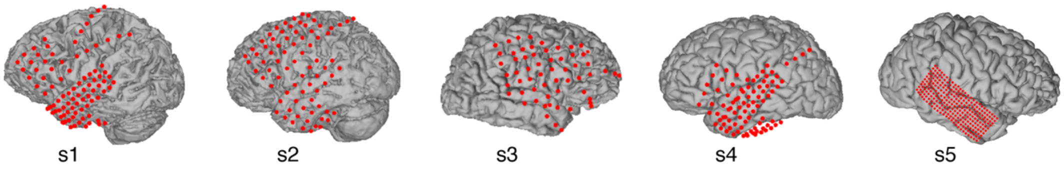


Figure 4-1 Electrode locations. *Grid locations for each subject were overlaid on cortical surface reconstructions of each subject's MRI scan.*

ECoG signals were recorded at the bedside using seven 16-channel g.USBamp biosignal acquisition devices (g.tec, Graz, Austria) at a sampling rate of 9,600 Hz. Electrode contacts distant from epileptic foci and areas of interest were used for reference and ground. Data acquisition and synchronization with the task presentation were accomplished using BCI2000 software (Schalk et al. 2004; Schalk 2010). All electrodes were subsequently downsampled to 1,000 Hz, corrected for DC shifts, and band pass filtered from 0.5 to 200 Hz. Notch filters at 60 Hz, 120 Hz and 180 Hz were used to remove electromagnetic noise. The time series were then visually inspected to remove the intervals containing ictal activity as well as electrodes that had excessive noise (including broadband electromagnetic noise from hospital equipment or poor contact with the cortical surface). Finally, electrodes were re-referenced to a common average. Imagined speech trials were carefully analyzed to remove those that were contaminated by overt speech. Overt speech trials that had grammar mistakes were also removed.

In addition to the ECoG signals, we acquired the subject's voice through a dynamic microphone (Samson R21s) that was rated for voice recordings (bandwidth 80-12,000 Hz, sensitivity 2.24 mV/Pa) and placed within 10 cm of the patient's face. We used a dedicated 16-channel g.USBamp to amplify and digitize the microphone signal in sync with the ECoG data. Finally, we verified the patient's compliance in the imagined task using an eye-tracker (Tobii T60, Tobii Sweden).

4.3.2 Experimental paradigm

We used a word repetition task (overt and imagined) cued with an auditory stimulus presentation. Each trial started with an auditory cue presented through a loudspeaker indicating one of six individual words (average length = 800 ms ± 20) to repeat; 800 ms after the end of the auditory stimulus a cross was displayed on the screen for 1500 ms. This indicated to the subjects to imagine hearing the word again in their mind. Subjects were instructed to "imagine hearing", because we were interested in the auditory perceptual representation induced by imagery, rather than kinesthetic (imagine saying words) or visual (imagine seeing words) representations. Finally, after 500 ms of blank screen, a second cross was displayed for 1500 ms, and subjects had to repeat the word out loud. The choice of stimuli was carefully chosen to maximize variability in terms of acoustic features, number of syllables and semantic categories, but minimize word length variability (variance in word length 20 ms; 'spoon', 'cowboy', 'battlefield', 'swimming', 'python', 'telephone'). Trials were repeated randomly between 18 and 24 times. The precise task design and timing is summarized in *Figure 4-2*. The microphone recording was used to verify that subjects were not producing audible speech during imagery, as well as monitoring the behavior (speech onset and word length) during overt speech. For each condition, we analyzed high gamma activity (HG) and built separate, independent classifiers, which allowed us to compare classification accuracy and discriminative information across perception and imagery tasks.

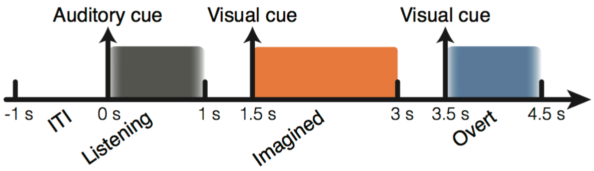


Figure 4-2 Experimental paradigm. *Subjects were presented with an auditory stimulus that indicated one of six individual words (average length = 800 ms ± 20). Then, a cue appeared on the screen [describe what the cue is and where it appeared on the screen], and subjects had to imagined hearing the word they had just listened to. Finally, a second cue appeared, and subjects had to say the word out loud. Shaded areas represent the intervals extracted for classification. For both listening and overt speech condition, we extracted epochs from 100 ms before speech onset to 100 ms after speech offset. For the imagined speech condition, we extracted fixed length 1.5sec epochs starting at cue onset, since there was no speech output.*

4.3.3 Feature extraction

To generate input features for the classifier, we filtered the ECoG signal in the high gamma (HG) frequency band (70-150 Hz; hamming window non-causal filter of order 20), and extracted the envelope using the Hilbert transform. Prior to model fitting, we downsampled the HG signal to 100 Hz to reduce computational load. For both listening and overt speech condition, we extracted epochs from 100 ms before speech onset to 100 ms after speech offset (unless otherwise stated). For the imagined speech condition, because there was no speech output we extracted fixed 1.5 s epochs starting at cue onset.

4.3.4 Classification

To classify the different pairs of words, we used support-vector machines (Hastie 2009) (SVM). This classifier maps the original input features into a higher dimensionality non-linear feature space via a kernel function. Its main advantages are robustness to overfitting (due to the inclusion of a regularization term) and underfitting (Stanikov et al. 2011) (due to the higher-dimensional mapping of the features). The general approach of SVM is described as follows:

$$F(x) = \mathrm{sgn}\left(\sum_{i} \alpha_i t_i K(x, x_i) + \beta_0\right)$$

where $x \in \mathbb{R}^P$, with P the number of features. K is the kernel function that transforms the input data x into a non-linear feature map. $\alpha_i \in [0, C]$ are weights for the support vectors. The constant C is the soft margin parameter, and controls the trade-off between classification error on the training set and smoothness of the decision boundary. t_i is the label of sample i, and β_0 is the offset of the separating hyperplane from the origin. For all the computations, we used the LIBSVM package (Chang and Lin 2011).

4.3.5 SVM-kernel computation

Traditionally, the Gaussian kernel is a widely used function used in SVM-classification. In this approach, the output of the classifier is based on a weighted linear combination of similarity measures (i.e., Euclidean distance) – computed between a data point and each of the support vectors (Hastie 2009). Speech production (overt and imagined) is subject to temporal variability (speech onset delays and local stretching/compression) across repetitions of the same instance. A classifier that assumes fixed time features might not recognize two trials as belonging to the same class if the neural patterns are not aligned in time. In order to deal with speech temporal irregularities, we developed a classification approach that incorporated non-linear time alignment in the kernel computation, using dynamic time warping (Sakoe and Chiba 1978; Rabiner 1993) (DTW; see section “Dynamic time warping” for details). The use of DTW-distances as an SVM-kernel function has shown its superiority over hidden Markov models for speech recognition. DTW provides a distance between two realigned time series that reflects how similar both are when maximally aligned (*Figure 4-3 B-C*). For each electrode separately, we computed the DTW-distance between each pair of trials (*Figure 4-3 A-C*). This gave rise to one kernel matrix per electrode (*Figure 4-3 D*). For the final kernel computation, we used a multiple kernel learning approach (Gönen and Ethem 2011) (MKL; see section “Multiple kernel learning” for details) to deal with the multiple kernel matrices – by doing a weighted average of the kernels associated with each electrode (*Figure 4-3 E*). The weighting was based on the discriminative power index of each individual electrode, which quantified the difference between the “within” class versus “between” class distances distribution (see section “Discriminative power index” for details).

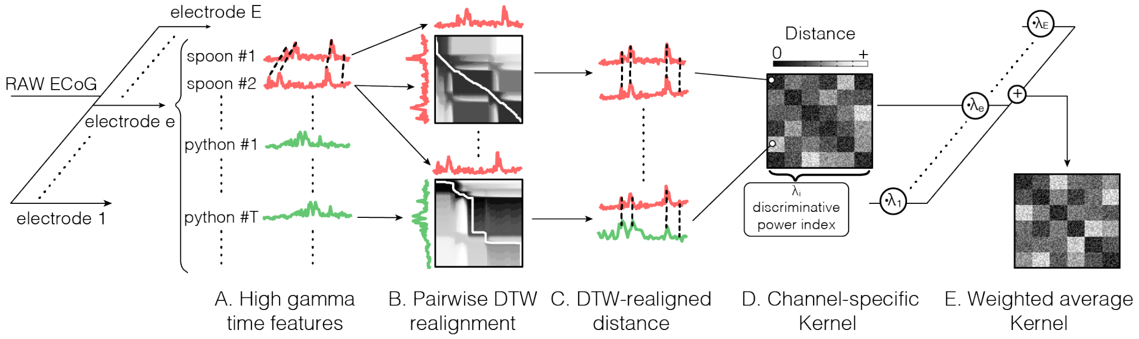

electrode E
RAW ECoG
electrode e
electrode 1
spoon #1
spoon #2
python #1
python #T
Distance
0
+
λi
discriminative power index
·λE
·λe
·λ1
+
A. High gamma time features
B. Pairwise DTW realignment
C. DTW-realigned distance
D. Channel-specific Kernel
E. Weighted average Kernel

Figure 4-3 Neural time course alignment. ***(A)*** *For each electrode separately, we extracted the high gamma time features.* ***(B)*** *We used dynamic time warping to realign the time series of each pair of trials, and* ***(C)*** *computed the DTW-distance between the pairwise realigned trials.* ***(D)*** *This gave rise to one similarity matrix per electrode (channel-specific kernel) that reflects how similar trial-pairs are after realignment. From the similarity matrix in d, we computed the discriminative power index (see Materials and methods for details).* ***(E)*** *The final kernel was computed as the weighted average of the individual kernels over all electrodes, based on their discriminative power index.*

Dynamic time warping

The main idea behind DTW is to locally stretch or compress (i.e., warp) two time series $x \in \mathbb{R}^M$ and $y \in \mathbb{R}^N$ where M and N are the number of time samples in x and y respectively (Sakoe and Chiba 1978). For each electrode separately, we computed the DTW for each pair of trials as follows:

Let $x^e \in \mathbb{R}^M$ and $y^e \in \mathbb{R}^N$, be the temporal features associated with two trials from electrode e. Each trial corresponded to a single word in one condition, represented as its associated HG features (see previous subsection). Trials had different length for both listening and overt speech conditions (M≠N), but equal length for the imagined speech condition (M=N). First, a pattern matching matrix d was computed between each time point pairs (*Figure 4-3 B*), as follows:

$$d(m,n) = f(x^e_m, y^e_n)\,, \qquad d \in \mathbb{R}^{MxN}$$

where $d(m,n)$ is the pattern matching index between $x^e_m$and y^e_n at the time sample m and n, respectively, and f an arbitrary distance metric. In this study, we used the Euclidean distance defined as $d(m,n) = \sqrt{(x^e_m - y^e_m)^2}$. Given a warping path φ, the average accumulated distortion between both warped signals is defined by:

$$d_\varphi(x^e, y^e) = \frac{1}{K}\sum_{k=1}^{K} d\left(\varphi_x(k), \varphi_y\,(k)\right),$$

where φ_x and φ_y are the warping functions of length K (that remap the time indices of x^e and y^e, respectively). The optimal warping path φ (white line in *Figure 4-3 B*), chooses the indices of x^e and y^e in order to minimize the overall accumulated distance:

$$D_e(x,y) = \min_\varphi d_\varphi\,(x,y),$$

where D_e is the optimal realigned Euclidean distance between x and y at a given electrode e. A dynamic programming approach was used to solve the global distance efficiently (Ellis 2003).

Multiple kernel learning

Once the realigned DTW distances of each electrode were computed, we built the kernel for the SVM classification by summing the weighted DTW kernels (fixed-rule multiple kernel (Gönen and Ethem 2011), as follows:

$$K(x,y) = \sum_{e} \lambda_e K_e(x,y) = \sum_{e} \lambda_e \exp\left(-\frac{D_e(x,y)}{\gamma}\right)$$

where $\lambda_e \in [0,1]$ is the normalized discriminative power index of electrode e, $\sum_e \lambda_e = 1$; and $\gamma > 0$ a free parameter.

Among the many ways to compute discriminative power between classes, we opted for the area under the receiver operating characteristic curve (Hastie 2009) (ROC), which measured the performance of a linear discriminant trained on the features given by the kernel K_e. This index reflected the difference between the “within” class versus “between” class distances. Entries of the distance matrices representing the within class distance had label = 0 and entries of the distance matrix representing the between class distance had label = 1. For each electrode, the discriminative power index was calculated from the data with 0/1 labels and the corresponding realigned DTW distance matrices.

4.3.6 Evaluation

Due to the limited number of trials, the classification performance was evaluated using a leave-one-out cross validation, where the test set was composed of one sample per class for each fold. With the training data, an inner leave-one-out cross validation was performed to find the optimal free parameters γ and C using a grid search approach, and were then fixed for the test evaluation. We also computed the discriminative power index λ_e on the training set. Here, we reported the classification accuracy as the percentage of correctly classified trials averaged on the outer loop cross-validation.

4.3.7 Statistics

For each condition, we evaluated statistical significance for each pair of words using randomization tests. After extracting the high gamma time features, we randomly shuffled 1,000 times the trial labels, and applied the exact same approach as in the actual classification process; we extracted HG time features, split into training and testing set, applied DTW at the trial level, built the kernel function, performed grid search on the inner loop, built the final model and evaluated the accuracy on the outer loop testing set. The averaged accuracy across cross-validated testing set yielded one value in the null distribution. The proportion of shuffled classification accuracy values greater than the observed accuracy yielded the p-value that the observed accuracy was due to chance. We corrected for multiple comparison using False Discovery Rate (Benjamini and Hochberg 1995) (FDR). The average of the null distributions was not significantly different from the expected value of chance level (50%; p>0.5; one-sample t-test). For each individual subject, we evaluated the significance level of the classification accuracy across all word pairs using one-sample t-tests against chance level (50%). Here again, we corrected for multiple comparison using FDR. Finally, we evaluated the significance level of the average classification accuracy across subjects using one-sample t-test again chance level. To investigate possible anatomical differences between conditions, all electrodes carrying significant discriminative information in at least one condition (listening, overt or imagined; $p < 0.05$; Bonferroni correction) were selected for an unbalanced Two-Way ANOVA with interactions, with experimental condition (listening, overt and imagined) and anatomical region (superior temporal gyrus, inferior frontal gyrus and sensorimotor cortex) as factors. Prior to ANOVA, we performed Mauchly's test to ensure the sphericity of data (Mauchly 1940). To define the significance level of discriminative power of each single electrode, we computed the discriminative power indices for the above-mentioned shuffled data. The discriminative power index yielded one value in the null distribution. For each electrode, the proportion of shuffled index values greater than the observed value yielded the p-value that the observed discriminative power of the electrode was due to chance.

4.4 Results

4.4.1 High gamma features

We analyzed z-scored high gamma time courses at different electrode locations, and compared the different conditions (listening, overt, and imagined speech). Word perception and production (both overt and imagined) evoked different high-gamma neural responses across many electrodes (*Figure 4-4A*) in all participants (Supplementary Figure 1). An exemplary electrode in the posterior superior temporal gyrus showed activation during all three conditions, while the neighboring electrode had activity only in the listening and overt speech conditions. An electrode in sensorimotor cortex showed sustained activity in the overt and imagined speech task. Finally, an electrode in the anterior temporal lobe, associated with speech production, exhibited activity only in the overt speech task, but not during listening or imagined speech. These results revealed the complex dynamics of speech perception and production (overt and imagined), and suggest that the neural representations underlying the different speech modalities are partially overlapping, yet dissociable (Yetkin et al. 1995; Rosen et al. 2000; Palmer et al. 2001).

In the listening condition, the auditory stimuli were time-locked across repetitions (*Figure 4-4 B*; audio envelope in red; standard deviation of the onset delay averaged over all words = 0 ms). Alternatively, in the overt speech condition, temporal irregularities in the speech onset and word duration were observed across repetitions of the same utterance (*Figure 4-4 B*; standard deviation of the onset delay averaged over all words = 220 ms). Because high gamma neural activity is known to track the speech envelope (Pasley et al. 2012; Mesgarani and Chang 2012; Martin et al. 2014; Kubanek et al. 2013), we assumed that temporal variations in overt and imagined speech would also be represented by the measured neural responses. As such, a classifier that assumes fixed neural temporal features would not recognize two trials as belonging to the same class if the neural patterns are not aligned in time. To overcome this limitation, we applied a temporal realignment procedure in the feature map generation. The procedure was applied to both overt and imagined speech, as both conditions have been shown to be subject to similar speech production temporal variations (Hubbard 2010).

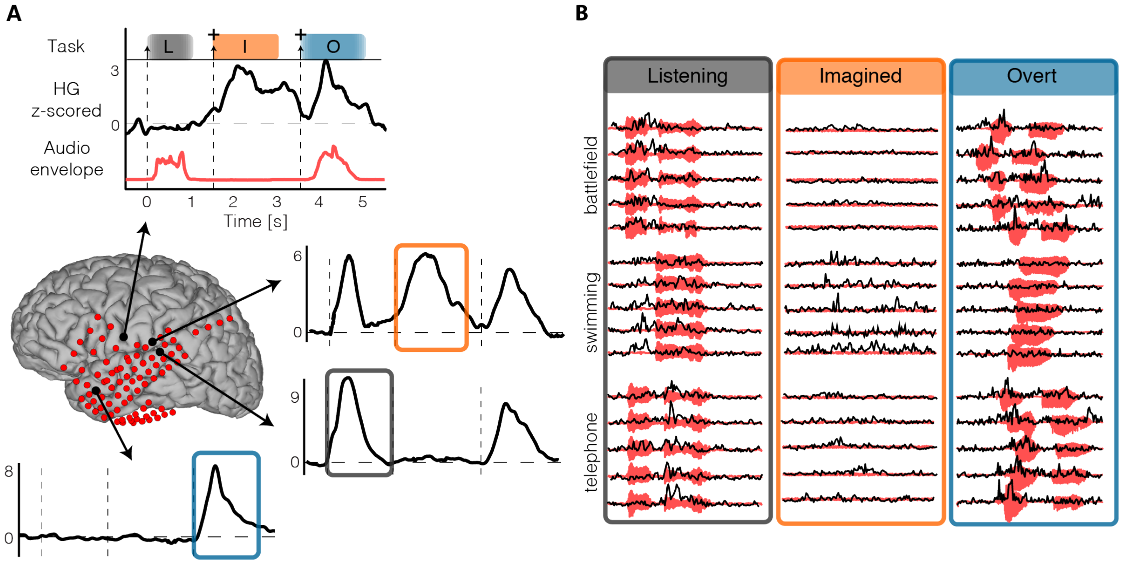


Figure 4-4 High gamma time course. ***(A)*** *High gamma neural activity averaged across trials and z-scored with respect to the pre-auditory stimuli baseline condition (500 ms interval). The top-most plot displays the designed task, an example of averaged time course for a representative electrode and the averaged audio envelope (red line).* ***(B)*** *For the given electrodes and conditions (listening, imagined and overt speech), examples of individual trials (black) and their corresponding audio recording (red) for three different words ('battlefield', 'swimming' and 'telephone').*

4.4.2 Classification

We used support-vector machines (Hastie 2009) to perform pair-wise classification of different individual words in the three different speech conditions (overt, listening and imagined speech). We first extracted the high gamma using bandpass filtering in the 70-150 Hz range, extracted the envelope using the Hilbert transform. We then extracted epochs from 100 ms before speech onset to 100 ms after speech offset for both listening and overt speech condition. Average word length for both listening and overt speech conditions were 800 ms ± 20 and 766 ms ± 84, respectively. In the imagined speech condition, due to the lack of speech output, we extracted 1500 ms epochs starting at cue onset (see Materials and methods for details).

In the imagined speech condition, pairwise classification accuracy reached 88.3% for one classification pair in a subject with extensive left temporal coverage (subject 4; *Figure 4-5 A*). Eight out of fifteen word-pairs were classified significantly higher than chance level ($p < 0.05$; randomization test; FDR correction), exceeding the number of pairs expected by chance ($0.05 \times 15 = 0.75$). As expected, higher classification accuracy was obtained in the listening and overt speech conditions where speech stimuli were directly observed. For both conditions, pairwise classification accuracy approached 100% in some comparisons, and twelve and fifteen out of fifteen pairs were significantly above chance, respectively ($p < 0.05$; randomization test; FDR correction).

Classification accuracy varied across subjects and pairs of words. In 4 out of 5 subjects, classification accuracy over all word pairs was significant in the imagined speech condition (*Figure 4-5 B*; $p < 0.05$; one-sample t-test; FDR correction), while the last subject was not significantly better than chance level (mean = 49.8%; $p > 0.5$; one-sample t-test; FDR correction). For listening and overt speech conditions, classification accuracy over all word pairs was again significant in all four subjects, and ranged between 83.0% and 96.0% ($p < 10^{-4}$; one-sample t-test; FDR correction).

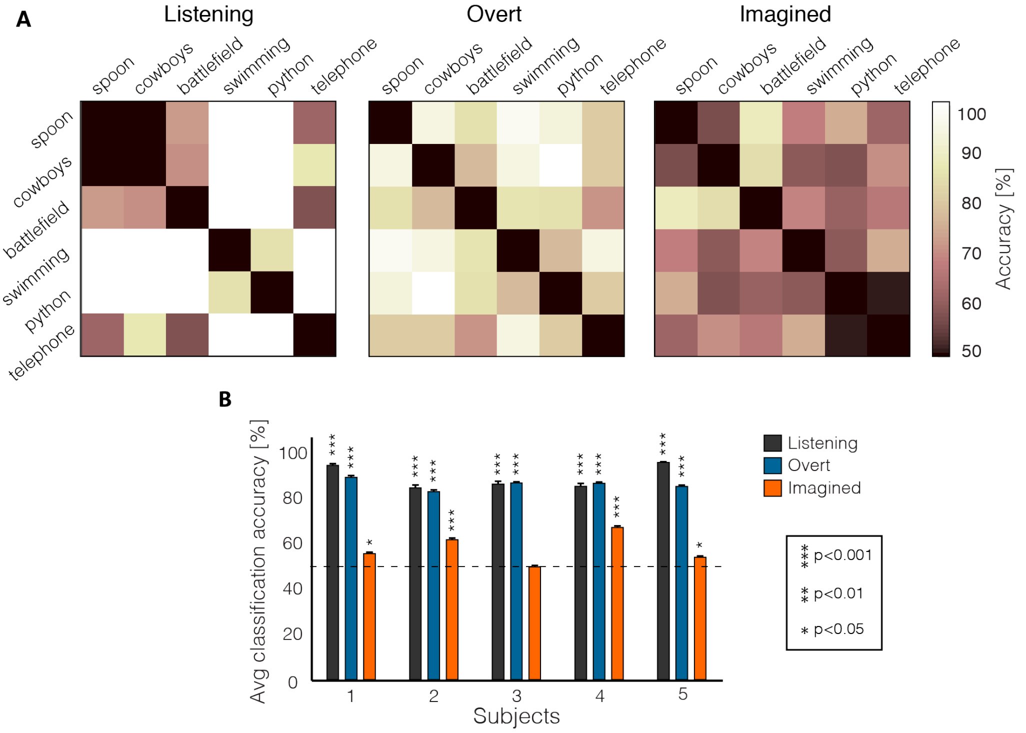


Figure 4-5 Classification accuracy. ***(A)*** *Pairwise classification accuracy in the testing set for the listening (left panel), overt speech (middle panel) and imagined speech condition (right panel) for a subject with good temporal coverage (S4).* ***(B)*** *Average classification accuracy across all pairs of words for each subject and condition (listening, overt and imagined speech). Error bars denote SEM.*

At the population level, average classification accuracy across all pairs was above chance level in all three conditions (*Figure 4-5 B*; listening: mean = 89.4% $p < 10^{-4}$; overt speech: mean = 86.2%, $p < 10^{-5}$; imagined speech: mean = 57.7%; $p < 0.05$; one-sample t-tests; FDR correction). A repeated measure 1-way ANOVA with experimental condition as a factor confirmed a difference among conditions ($F_{(2, 12)} = 56.3$, $p < 10^{-5}$). Post-hoc t-tests showed that the mean classification accuracy for listening was not significantly different from the overt speech ($p > 0.1$; two-sample t-test; FDR correction). Both were significantly higher than the imagined speech classification accuracy ($p < 0.005$; two-sample t-test; FDR correction). Although the classification accuracy for imagined speech was lower than for listening and overt speech, the imagery classification results provide evidence that high gamma time course during imagined speech contained information to distinguish pairs of words.

To assess the impact of the neural activity realignment procedure in classification accuracy, we evaluated the improvement of DTW alignment compared to when no alignment was applied. The results showed that for both the overt and imagined speech conditions, the average classification accuracy was y reduced when no alignment was applied (*Supplementary Figure 2 A*; $p < 0.05$; two-sample t-test; FDR correction). On the other hand, for the listening condition – in which trials were time-locked to stimulus onset – the DTW procedure did not improve the classification accuracy ($p> 0.5$; two-sample t-test; FDR correction).

The inability to directly measure temporal variability in the imagery condition remains a major limiting factor for classification accuracy, despite the realignment procedure we employed. In the imagined speech condition, due to the lack of speech output, we could only extract trials at cue onset rather than at the true onset of speech imagery. To investigate the impact of this limitation on classification accuracy, we analyzed data from the overt speech condition where the auditory stimulus is directly measured. The results showed that classification accuracy in the overt speech condition was reduced when extracting epochs at cue onset, compared to when epochs were extracted between speech onset and offset (*Supplementary Figure 2 B*; $p < 0.05$; two-sample t-test). This further highlights limitations in the realignment algorithm, and indicates that imagery classification accuracy may be increased by developing enhanced methods to define imagined speech onset and offset.

4.4.3 Anatomical distribution of discriminant electrodes

To assess how the brain areas important for word classification vary across experimental conditions, we analyzed the anatomical distribution of the electrodes carrying discriminative information in the three different conditions. For each electrode and condition, we computed a discriminative power index that reflected the predictive power of each electrode in the classification process (see Materials and methods for details).

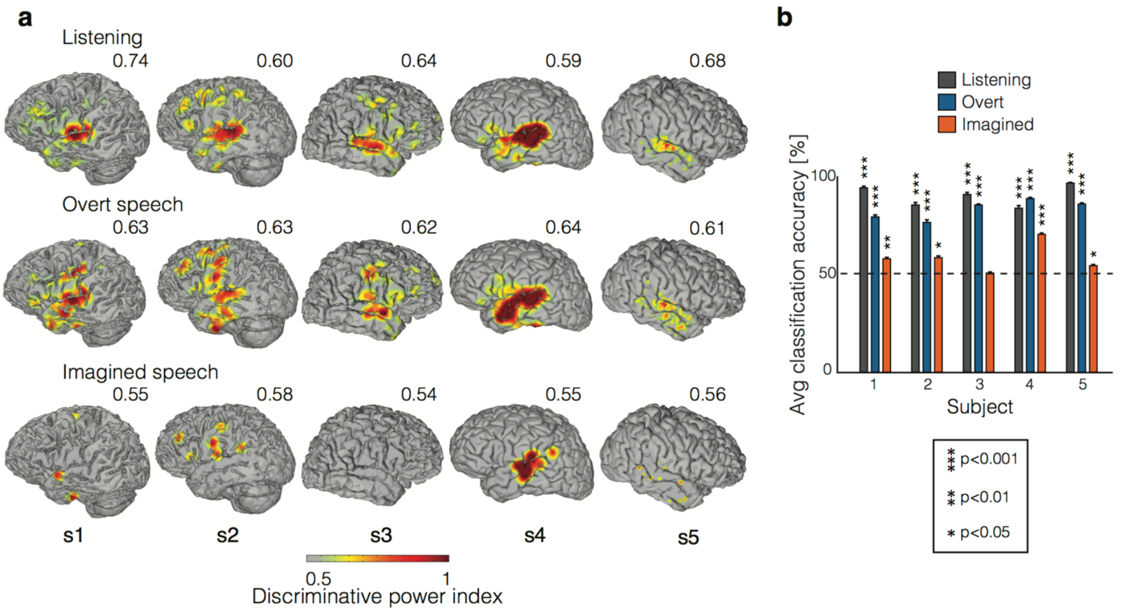


Figure 4-6 Discriminative information. (A) *Discriminative power measured as the areas under the ROC curve (thresholded at $p < 0.05$; uncorrected; see Materials and methods for details), and plotted on each individual's brain. Each is scaled to the maximum absolute value of discriminative power index (indicated by the number above each cortical map).* ***(B)*** *Average classification accuracy across all pairs of words for each subject using only temporal electrodes for the listening (top panel), overt speech (middle panel) and imagined speech (bottom panel). Error bars denote SEM.*

Figure 4-6 A shows the anatomical distribution of the discriminative power index across each condition (heat map thresholded at $p < 0.05$; uncorrected). Overall, the highest discriminative information was located in the temporal gyrus, inferior frontal gyrus and sensorimotor gyrus – regions commonly associated with speech processing. Anatomical differences between conditions were assessed for significant electrodes (188 electrodes significant in at least one condition; $p < 0.05$; FDR correction), using an unbalanced Two-Way ANOVA with interactions, with experimental condition (listening, overt and imagined speech) and anatomical region (superior temporal gyrus (STG), inferior frontal gyrus (IFG) and sensorimotor cortex (SMC)) as factors. The main effect of experimental condition was significant [$F_{(2, 555)} = 29.1$, $p < 10^{-15}$], indicating that the discriminative

information in the classification process was different across conditions. Post-hoc t-tests with Bonferroni correction showed that the overall discriminative power was higher in the listening (mean = 0.56) and overt speech condition (mean = 0.56) than in the imagined speech (mean = 0.53; $p < 10^{-10}$; unpaired two-sample t-test; Bonferroni correction), at the level of single electrodes. The main effect of anatomical region was also significant [$F_{(2, 555)} = 7.18$, $p < 0.001$]. Post-hoc t-tests indicated stronger discriminative information in the STG (mean = 0.55) than in the inferior frontal gyrus (mean = 0.54; $p < 0.05$; unpaired two-sample t-test; Bonferroni correction), but not than the SMC (mean = 0.54; $p > 0.05$; unpaired two-sample t-test; Bonferroni correction). The interaction between gyrus and experimental condition was also significant [$F_{(4, 555)} = 6.7$; $p < 10^{-4}$]. Specifically, The discriminative power in the STG was higher for listening (mean = 0.57) and overt speech (mean = 0.56) than for imagined speech (mean = 0.53; $p < 10^{-10}$; unpaired two-sample t-test; Bonferroni correction). In addition, the discriminative power in the sensorimotor cortex was higher in the overt condition (mean = 0.57), than in the listening (mean = 0.54) and imagined condition (mean = 0.53; $p < 0.001$; unpaired two-sample t-test; Bonferroni correction). Similarly, the frontal electrodes provided more discriminative information in the overt speech (mean = 0.55) than in the imagined speech condition (mean = 0.53; $p < 10^{-4}$; unpaired two-sample t-test; Bonferroni correction). Post-hoc t-tests also showed that the discriminative power in the listening condition was higher in the STG (mean = 0.56) than in the IFG (mean = 0.54) and SMC (mean = 0.54; $p < 0.05$; unpaired two-sample t-test; FDR correction). Finally, no significant differences across gyri were observed in the imagined speech condition ($p > 0.5$; unpaired two-sample t-test; Bonferroni correction).

While the anatomic locations (i.e. STG, IFG and SMC) that give rise to the best word discrimination in the listening and overt speech conditions were consistent across subjects, discriminative anatomic locations in the imagined condition varied. To further investigate brain areas and based on a number of previous studies demonstrating its role in auditory imagery (Yetkin et al. 1995; McGuire et al. 1996; Palmer et al. 2001; Shergill et al. 2001; Pei et al. 2011), we performed the classification using only electrodes from the superior temporal gyrus (*Figure 4-6 B*). In the imagery condition, classification accuracy using STG electrodes was significant in four out of five subjects ($p < 0.05$; one-sample t-test; FDR correction), while it was not significant in S3 ($p > 0.5$; one-sample t-test; FDR correction). At the group level, classification using only temporal electrodes was significant (mean = 58.0%; $p < 0.05$; one-sample t-test; FDR correction). For both, listening and overt speech conditions, classification accuracy was significant in all individual subjects when using only STG electrodes ($p < 10^{-4}$; one-sample t-test; FDR correction), as well as at the group level (mean listening = 89.5% and mean overt speech = 82.6%; $p < 10^{-4}$; one-sample t-test; FDR correction). This provides preliminary evidence that superior temporal gyrus alone could drive auditory imagery decoding, but that other areas such as frontal cortex and sensorimotor cortex could also contribute.

4.5 Discussion

Our results provide the first demonstration of single-trial neural decoding of words during imagined speech production. We developed a new binary classification approach that accounted for temporal variations in the high gamma neural activity across speech utterances. We used support-vector machines to classify individual words in a word pair, and introduced a non-linear time alignment into the kernel to deal with internal speech production variability. At the group level, average classification accuracy across all pairs was significant in all three conditions. Two subjects that exhibited the lowest classification scores had right hemisphere coverage and were right handed, which is typically associated with left hemisphere language dominance (Toga and Thompson 2003). This may have contributed to differences in accuracy across left and right hemisphere grid subjects. However, more data are required to delineate the effect of hemisphere coverage in the decoding process. The anatomic locations that led to the best word discrimination in the listening and overt speech conditions were consistent across subjects. All three anatomical regions (STG, IFG and SMC) provided information in the classification process. In the imagery condition, anatomical areas with the highest predictive power were more variable across subjects. The results revealed that the STG alone could drive auditory imagery decoding, but that other areas, such as the IFG and SMC also contribute.

An important component of the study is the application of dynamic time warping in the classification framework to account for speech production temporal irregularities. This technique maximizes alignment of the neural activity time courses without knowledge of the exact onset of the events. This approach proved useful for studying imagined speech where no behavior or stimuli are explicitly observed. In contrast, DTW did not improve accuracy in the listening condition, where neural activity is already time-locked to stimulus events. This highlights the usefulness of a time alignment procedure such as using DTW for modeling the neural activity of unobserved behavioral events such as imagery. We also note the limitations of DTW in noisy environments suggesting that imagery results may be improved by developing more robust realignment techniques. We also show that overt speech classification accuracy was improved when epochs were selected from speech onset/offset, as compared to when they were extracted from cue onset. This suggests that the results may be improved by developing enhanced methods to define imagined speech onset and offset. Ideas for possible future directions would be to improve experimental paradigms (i.e. button press, karaoke-task, etc.), define improved behavioral or neural metrics that correlates with speech onset/offset and increased training in imagery prior to ECOG recording.

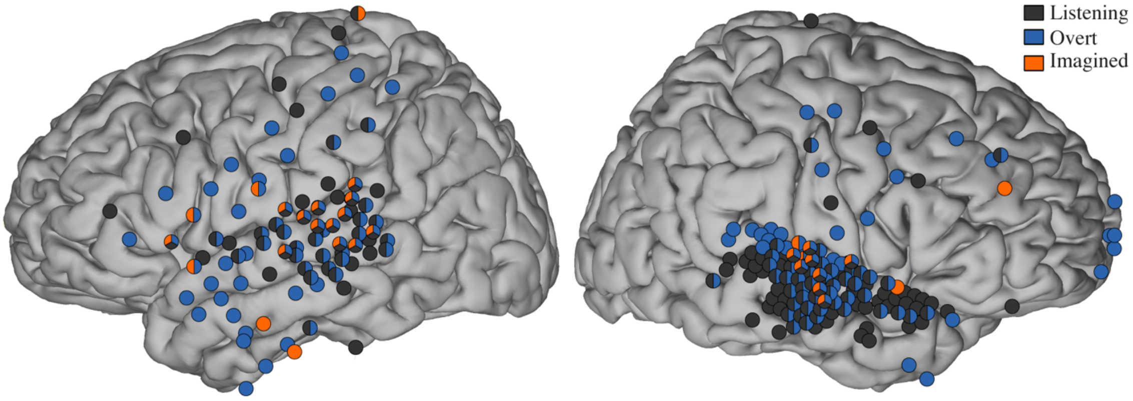
Despite intense investigation, it is still unclear how the content of imagined speech is processed in the human cortex. Different tasks – such as word repetition, letter or object naming, verb generation, reading, rhyme judgment, counting – involve different speech production processes, ranging from lexical retrieval to phonological or even phonetic encoding (Perrone-Bertolotti et al. 2014). In this study, we chose the set of auditory stimuli to maximize variability in several speech feature spaces (acoustic features, number of syllable, semantic categories), but to minimize word length variance. Our approach does not allow us to investigate which specific speech features provided information and allowed classification; i.e., if the discrimination was based on acoustic, phonetic, phonological, semantic or abstract features within speech perception, comprehension or production. Given that several brain areas were involved, it is likely that various features of speech were involved in the classification process.

Several additional limitations precluded high word prediction accuracy during imagined speech. First, we were limited by the electrode location and duration of implantation that was not designed for the experiments, but solely for clinical needs. Higher density grids placed at specific locations in the posterior superior temporal gyrus, frontal cortex and/or sensorimotor cortex that are active during imagined speech would provide higher spatial resolution and potentially enhanced discriminating signals. Further, subjects were not familiarized with the task beforehand (i.e. no training), and due to time constraints in the epilepsy-monitoring unit, we were unable to monitor subjects' performance or vividness during speech imagery. We also could not reject pronunciation and grammatical mistakes, as we did in the overt speech condition. We propose it would be beneficial to train subjects on speech imagery prior to surgery to enhance task performance.

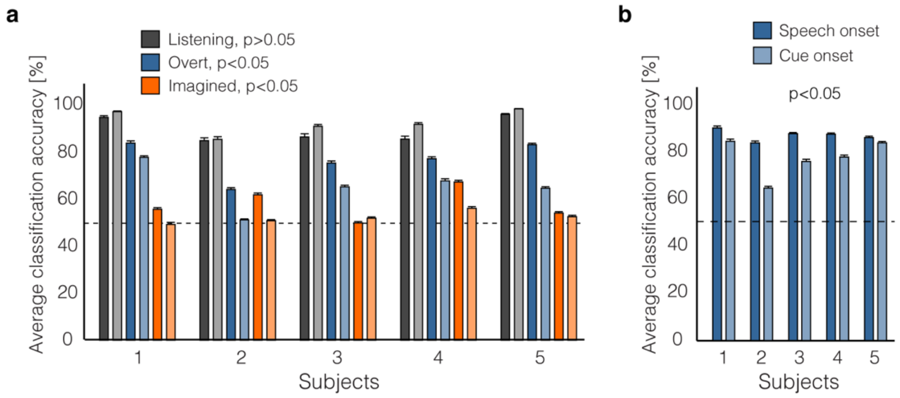
Finally, although our study is a proof of concepts for basic decoding of speech imagery, many issues still need to be tackled to prove the feasibility for a clinical application. Our current approach was limited in the set of choices available, and only tests binary classification between word pairs. In addition, the effect size is small, and likely not clinically significant for a communication interface.

Classification of individual words among multiple other words or continuous speech decoding would be a more realistic clinical scenario. An alternative would be classifying phonemes, which forms the building blocks of speech instances. Decoding vowels and consonants in overt and imagined words using electrocorticographic signals in humans has shown promising results (Brumberg et al. 2011;'Herff et al. 2015), and would allow generating a larger lexicon from a fewer number of classes (60-80 phonemes in spoken English (Vaseghi 2007)).

4.6 Supplementary Material



Supplementary Figure 1. Active electrodes across all subjects. *We computed the average amplitude from the time course described in Figure 4-4 A for all three conditions. We computed the coefficient of determination (r^2) between baseline and active state (listening, overt and imagined speech) for each electrodes(Wonnacott 1990). To define the significance level of each electrode, we shuffled the labels 1,000 times, and computed r^2. The proportion of shuffled r^2 greater than the observed r^2 yields the p-value that the observed activation is due to chance. Electrodes with p<0.05 (corrected for multiple comparison with False Discovery Rate) were plotted on the Talairach brain.*



Supplementary Figure 2. Effect of dynamic time warping realignment. ***(A)*** *Average classification accuracy across all pairs of words for each subject and condition (listening, overt and imagined speech) – using DTW (dark) and without DTW (light). Error bars denote resampling SEM.* ***(B)*** *Comparison of the classification* *accuracy in the overt speech condition – when epochs were extracted at speech onset or at cue onset. Error bars denote SEM.*

Chapter 5 Neural encoding of auditory features during music imagery

Disclaimer: This chapter is adapted from the following article – with permissions of all co-authors and journals:

Martin S., Mikutta C., Leonard M.K., Hungate D., Koelsch S., Chang E.F., Millán J.d.R., Knight R.T., Pasley B.N. 2017. "Neural encoding of auditory features during music perception and imagery. (under revision at Cerebral Cortex). *bioRxiv* doi:10.1101/106617.

My contribution: Conceptualization, formal Analysis, methodology, visualization, writing – original draft preparation.

5.1 Abstract

It remains unclear how the human cortex represents spectrotemporal sound features during auditory imagery, and how this representation compares to auditory perception. To assess this, we recorded electrocorticographic signals from an epileptic patient with proficient music ability in two conditions. First, the participant played two piano pieces on an electronic piano with the sound volume of the digital keyboard on. Second, the participant replayed the same piano pieces, but without auditory feedback, and the participant was asked to imagine hearing the music in his mind. In both conditions, the sound output of the keyboard was recorded, thus allowing precise time-locking between the neural activity and the spectrotemporal content of the music imagery. For both conditions, we built encoding models to predict high gamma neural activity (70-150Hz) from the spectrogram representation of the recorded sound. We found robust similarities between perception and imagery – in frequency and temporal tuning properties in auditory areas.

5.2 Introduction

Auditory imagery is defined here as the mental representation of sound perception in the absence of external auditory stimulation. The experience of auditory imagery is common, such as when a song runs continually through someone's mind. On an advanced level, professional musicians are able to imagine a piece of music by looking at the sheet music (Meister et al. 2004). Behavioral studies have shown that structural and temporal properties of auditory features (see (Hubbard 2010) for complete review), such as pitch (Halpern 1989b), timbre (Halpern et al. 2004; Pitt and Crowder 1992), loudness (Intons-Peterson 1992) and rhythm (Halpern 1988) are preserved during auditory imagery. However, it is unclear how these auditory features are represented in the brain during imagery. Experimental investigation is difficult due to the lack of observable stimulus or behavioral markers during auditory imagery. Using a novel experimental paradigm to synchronize auditory imagery events to neural activity, we investigated the neural representation of spectrotemporal auditory features during auditory imagery in an epileptic patient with proficient music abilities.

Previous studies have identified anatomical regions active during auditory imagery (Kosslyn, Ganis, and Thompson 2001), and how they compare to actual auditory perception. For instance, lesion

(Zatorre and Halpern 1993) and brain imaging studies (Griffiths 1999; Halpern 1999; Halpern et al. 2004; Kraemer et al. 2005; Rauschecker 2001; Zatorre et al. 1996) have confirmed the involvement of bilateral temporal lobe regions during auditory imagery (see (Zatorre and Halpern 2005) for a review). Brain areas consistently activated with fMRI during auditory imagery include the secondary auditory cortex (Griffiths 1999; Kraemer et al. 2005; Zatorre, Halpern, and Bouffard 2009), the frontal cortex (Halpern and Zatorre 1999; Zatorre, Halpern, and Bouffard 2009), the sylvian parietal temporal area (Hickok et al. 2003), ventrolateral and dorsolateral cortices (Meyer et al. 2007) and the supplementary motor area (Halpern and Zatorre 1999; Halpern 2001; Mikumo 1994; Petsche, von Stein, and Filz 1996; Brodsky et al. 2003; Schürmann et al. 2002). Anatomical regions active during auditory imagery have been compared to actual auditory perception to understand the interactions between externally and internally driven cortical processes. Several studies showed that auditory imagery has substantial, but not complete overlap in brain areas with music perception (Kosslyn, Ganis, and Thompson 2001) – e.g. the secondary auditory cortex is consistently activated during music imagery and perception while the primary auditory areas appear to be activated solely during auditory perception (Bunzeck et al. 2005; Griffiths 1999; Halpern et al. 2004; Yoo, Lee, and Choi 2001).

These studies have helped to unravel anatomical brain areas involved in auditory perception and imagery. However, there is lack of evidence for the representation of specific acoustic features in the human cortex during auditory imagery. It remains a challenge to investigate neural processing during internal subjective experience like music imagery, due to the difficulty in time-locking brain activity to a measurable stimulus during auditory imagery. To address this issue, we recorded electrocorticographic neural signals (ECoG) of a proficient piano player in a novel task design that permitted robust marking of the spectrotemporal content of the intended music imagery to neural activity – thus allowing us to investigate specific auditory features during auditory imagery. In the first condition, the participant played an electronic piano with the sound output turned on. In this condition, the sound was played out loud through speakers at a comfortable sound volume that allowed auditory feedback (perception condition). In the second condition, the participant played the electronic piano with the speakers turned off, and instead imagined the corresponding music in his mind (imagery condition). In both conditions, the digitized sound output of the MIDI-compatible sound module was recorded. This provided a measurable record of the content and timing of the participant's music imagery when the speakers of the keyboard were turned off and he did not hear the music. This task design allowed precise temporal alignment between the recorded neural activity and spectrogram representations of music perception and imagery – providing a unique opportunity to apply receptive field modeling techniques to quantitatively study neural encoding during auditory imagery.

A well-established role of the early auditory system is to decompose complex sounds into their component frequencies (Aertsen, Olders, and Johannesma 1981; Eggermont, Aertsen, and Johannesma 1983; Tian 2004), giving rise to tonotopic maps in the auditory cortex *(*see Saenz and Langers 2014 for a review). Auditory perception has been extensively studied in animal models and humans using spectrotemporal receptive field (STRFs) analysis (Aertsen, Olders, and Johannesma 1981; Chi, Ru, and Shamma 2005; Clopton and Backoff 1991; Pasley et al. 2012; Theunissen, Sen, and Doupe 2000), which identifies the time-frequency stimulus features encoded by a neuron or population of neurons. STRFs are consistently observed during auditory perception tasks, but the existence of STRFs during auditory imagery is unclear due to the experimental challenges associated with synchronizing neural activity and the imagined stimulus. To characterize and compare the spectrotemporal tuning properties during auditory imagery and perception, we fitted two encoding models on data collected from the perception and imagery conditions. In this case, encoding models describe the linear mapping between a given auditory stimulus representation and its corresponding brain response. For instance, encoding models have revealed the neural tuning properties of various speech features, such as acoustic, phonetic and semantic representations (Pasley et al. 2012; Lotte et al. 2015; Mesgarani et al. 2014; Tankus, Fried, and Shoham 2012; Huth et al. 2016).

In this study, the neural representation of music perception and imagery was quantified by spectrotemporal receptive fields (STRFs) that predict high gamma (HG; 70-150Hz) neural activity. High gamma correlates with the spiking activity of the underlying neuronal ensemble (Lachaux et al. 2012b; Miller et al. 2007; Boonstra, Houweling, and Muskulus 2009) and reliably tracks speech and music features in auditory and motor cortex (Pasley et al. 2012; Towle et al. 2008; Llorens et al. 2011; Crone et al. 2001; Sturm et al. 2014). Results demonstrated the presence of robust spectrotemporal receptive fields during auditory imagery with extensive overlap in frequency tuning and cortical location compared to receptive fields measured during auditory perception. These results provide a quantitative characterization of the shared neural representation underlying auditory perception and the subjective experience of auditory imagery.

5.3 Material and methods

5.3.1 Participant and data acquisition

Electrocorticographic (ECoG) recording was obtained using subdural electrode arrays implanted in one patient undergoing neurosurgical procedures for epilepsy. The participant volunteered and gave his informed consent before testing. The experimental protocol was approved by the University of California, San Francisco and Berkeley Institutional Review Boards and Committees on Human Research. Electrode grids had center-to-center distance of 4 mm. Grid placement and duration of ECoG monitoring were based solely on the requirements of the clinical evaluation. Localization and co-registration of electrodes was performed using the structural MRI. Multi-electrode ECoG data were amplified and digitally recorded with sampling rate of 3,052 Hz. ECoG signals were re-referenced to a common average after removal of electrodes with epileptic artifacts or excessive noise (including broadband electromagnetic noise from hospital equipment or poor contact with the cortical surface). In addition to the ECoG signals, the audio output of the piano was recorded along with the multi-electrode ECoG data.

5.3.2 Experimental design

The recording session included two conditions. In the first condition, the participant played on an electronic piano with the sound turned on. That is, the music was played out loud through the speakers of the digital keyboard in the hospital room (volume at comfortable and natural sound level; *Figure 5-1 A*; perception condition). In the second condition, the participant played on the piano with the speakers turned off and instead imagined hearing the corresponding music in his mind (*Figure 5-1 B*; imagery condition). *Figure 5-1 B* illustrates that in both conditions, the digitized sound output of the MIDI sound module was recorded in synchrony with the ECoG data (even when the speakers were turned off and the participant did not hear the music). The two music pieces were Chopin's Prelude in C minor Op. 28 no. 20 and Bach's Prelude in C major (BWV 846), respectively.

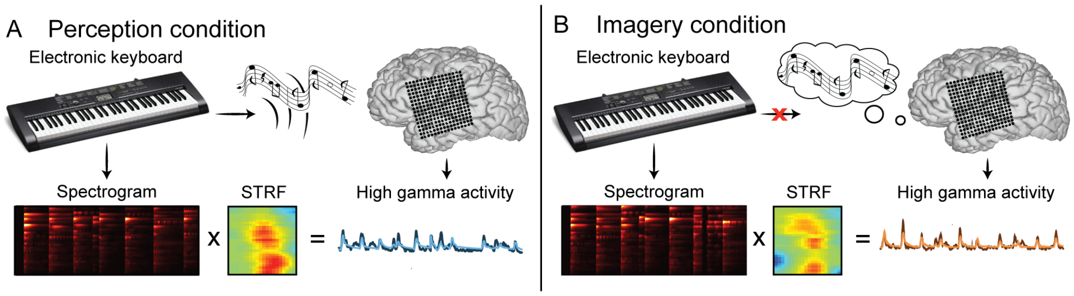


Figure 5-1 Experimental task design. ***(A)*** *The participant played an electronic piano with the sound of the digital keyboard turned on (perception condition).* ***(B)*** *In the second condition, the participant played the piano with the sound turned off and instead imagined the corresponding music in his mind (imagery condition). In both conditions, the digitized sound output of the MIDI-compatible sound module was recorded in synchrony with the neural signals (even when the participant did not hear any sound in the imagery condition). The models take as input a spectrogram consisting of time-varying spectral power across a range of acoustic frequencies (200– 7,000 Hz, bottom left) and output time-varying neural signals. To assess the encoding accuracy, the predicted neural signal (light lines) is compared to the original neural signal (dark lines).*

5.3.3 Feature extraction

We extracted the ECoG signal in the high gamma frequency band from eight bandpass filters (hamming window non-causal filter of order 20, logarithmically increasing center frequencies (70–150 Hz) and semi-logarithmically increasing bandwidths), and extracted the envelope using the Hilbert transform. The power was then calculated by averaging the signal across these eight bands. Subsequently, the signal was down-sampled to 100 Hz and z-scored.

5.3.4 Auditory spectrogram representation

We evaluated the ability to reconstruct the auditory spectrogram representation of music. The spectrogram is a time-varying representation of the amplitude envelope at 128 acoustic frequencies – logarithmically spaced between 180-7,000 Hz. This representation was generated by affine wavelet transforms of the sound pressure waveform using auditory filter banks mimicking the frequency analysis of the auditory periphery (Chi, Ru, and Shamma 2005). The spectrogram was subsequently downsampled to 32 frequency channels (unless otherwise stated). To compute these acoustic representations, we used the NSL MATLAB toolbox (http://www.isr.umd.edu/Labs/NSL/Software.htm).

5.3.5 Encoding model

The neural encoding model, based on the spectro-temporal receptive field (Theunissen, Sen, and Doupe 2000) describes the linear mapping between the music stimulus and the high gamma neural response at individual electrodes. The encoding model was estimated as follows:

$$\hat{R}(t,n) = \sum_{\tau}\sum_{f} h(\tau,f,n)S(t-\tau,f)$$

where $\hat{R}(t,n)$ is the predicted high gamma neural activity at time t and electrode n, $S(t-\tau,p)$ is the spectrogram representation at time $(t-\tau)$ and acoustic frequency f. Finally, $h(\tau,f,n)$ is the linear transformation matrix that depends on the time lag τ, the frequency f and electrodes n. h represents the spectro-temporal receptive field of each electrode. The STRFs are commonly used to estimate neural tuning to a wide variety of stimulus parameters in different sensory systems (Wu, David, and Gallant 2006).

We used Ridge regression to fit the encoding model (Thirion et al. 2011), and a 10-fold cross-validation resampling procedure, with no overlap between training and test partitions within each resample. We performed grid search on the training set to define the penalty coefficient α and the learning rate η, using a nested loop cross-validation approach. We standardized the parameter estimates to yield the final model.

5.3.6 Decoding model

The decoding model linearly mapped the neural activity to the music representation, as a weighted sum of activity at each electrode, as follows:

$$\hat{S}(t,f) = \sum_{\tau}\sum_{n} g(\tau,f,n)R(t-\tau,n)$$

where $R(t-\tau,n)$ is the high gamma neural response of electrode n at time $(t-\tau)$, where τ is the time lag ranging between -500 and 500ms. To reduce computational load, only electrodes that had significant forward predictions were taken into account for the decoding. $\hat{S}(t,p)$ is the estimated music representation at time t and frequency f, where f is one of 128 acoustic frequency features in the auditory spectrogram representation. Finally, $g(\tau,f,n)$ is the linear transformation matrix, which depends on the time lag τ, frequency f, and electrode, n. The music representation and the neural high gamma response data were synchronized, downsampled to 100 Hz, and standardized to zero mean and unit standard deviation prior to model fitting.

To fit model parameters, we used gradient descent with early stopping regularization. We used a 10-folds cross-validation resampling scheme, and standardized the parameter estimates to yield the final model. Within the training set, 20% of the data were used as validation set, to monitor out-of-sample prediction accuracy and determine the early stopping criterion. The algorithm was terminated after a series of 30 iterations failing to improve performance on the validation set or after 10,000 iterations. Finally, model prediction accuracy was evaluated on the independent testing set.

5.3.7 Statistical analysis

Prediction accuracy was quantified by computing the correlation coefficient (Pearson's r) between the predicted and actual HG signal using data from the independent test set for each fold and electrode. Overall encoding accuracy was reported as the mean correlation over folds. The z-test was applied for all reported mean r values. Electrodes were defined as significant if the p-value was smaller than the significance threshold of α=0.05 (95th-percentile; FDR correction). To define auditory sensory areas, we built an encoding model on data recorded while the participant listened passively to speech sentences from the TIMIT corpus (Garofolo 1993) during 10min. Electrodes with significant encoding accuracy are highlighted in *Supplementary Figure 5-1*.

To further investigate the neural encoding of spectrotemporal acoustic features during music perception and music imagery, we analyzed all the electrodes that were at least significant in one condition (unless otherwise stated). Tuning curves were estimated from spectro-temporal receptive fields, by first setting all inhibitory weights to zero, then averaging across the time dimension and converting to standardized z-scores. Tuning peaks were identified as significant peak parameters in the acoustic frequency tuning curves (z>3.1; p<0.001) – separated by more than one third an octave.

Decoding accuracy was quantified by computing the correlation coefficient (Pearson's r) between the reconstructed and original music representation using data from the independent test set. For each cross-validation resample, we calculated one correlation coefficient for each auditory feature over time – leading to 128 correlation coefficients for the auditory spectrogram representation. Overall reconstruction accuracy was reported as the mean correlation over resamples and speech components. Standard error of the mean (SEM) was calculated by taking the standard deviation of the overall reconstruction accuracy across resamples.

To further assess the predictive power of the reconstruction process, we evaluated the ability to identify specific blocks within the continuous recording. First, 0.5-second segments were extracted from the original and reconstructed spectrogram representations. Second, a confusion matrix was constructed where each element contained the similarity score between the target reconstructed segment and the original reference segments. To compute the similarity score between each target and reference segment, dynamic-time warping was applied to temporally align each pair and the mean correlation coefficient was used as the similarity score. The confusion matrix reflects how well a given reconstructed segment matches its corresponding original segment versus other candidates. The similarity scores were sorted, and identification accuracy was quantified as the percentile smaller than the rank of the correct segment. At chance level, the expected percentile rank is 0.5, while perfect identification is 1.0.

5.4 Results

5.4.1 High gamma neural encoding during auditory perception and imagery

Example auditory spectrograms from Chopin's Prelude determined through the participant's key presses with the electronic piano are shown in *Figure 5-2 B* for both perception and imagery conditions. To evaluate how consistently the participant performed across perception and imagery tasks, we computed the realigned Euclidean distance (Ellis 2003) between the spectrograms of the same music pieces played across conditions (within-stimulus distance). We compared the within-stimulus distance with the realigned Euclidean distance between the spectrograms of the different musical pieces (between-stimulus distance). The realigned Euclidean distance was 251.6% larger for the between-stimulus distance compared to the within-stimulus distance ($p<10^{-3}$; randomization test), suggesting that the spectrograms of same musical pieces played across conditions were more similar than the spectrograms of different musical pieces. This indicates that the participant performed the task with relative consistency and specificity across the two conditions. To compare spectrotemporal auditory representations during music perception and music imagery tasks, we fit separate encoding models in each condition. We used these models to quantify specific anatomical and neural tuning differences between auditory perception and imagery.

For both perception and imagery conditions, the observed and predicted high gamma neural responses are illustrated for two individual electrodes in the temporal lobe, respectively (*Figure 5-2 C),* together with the corresponding music spectrum (*Figure 5-2 B*). The predicted neural response for the electrode shown in the upper panel of *Figure 5-2 B* was significantly correlated with its corresponding measured neural response in both perception (r=0.41; $p<10^{-7}$; one-sample *z*-test; FDR correction) and imagery (r=0.42; $p<10^{-4}$; one-sample *z*-test; FDR correction) conditions. The predicted neural response for the lower panel electrode was correlated with the actual neural response only in the perception condition (r=0.23; p<0.005; one-sample *z*-test; FDR correction) but not in the imagery condition (r=-0.02; p>0.5; one-sample *z*-test; FDR correction). The difference between both conditions was significant for the electrode in the lower panel (p<0.05; two-sample t-test), but not in the upper panel (p>0.5; two-sample t-test). This suggests that there is a strong continuous relationship between time-varying imagined sound features and neural activity, but that this relationship is dependent on cortical location.

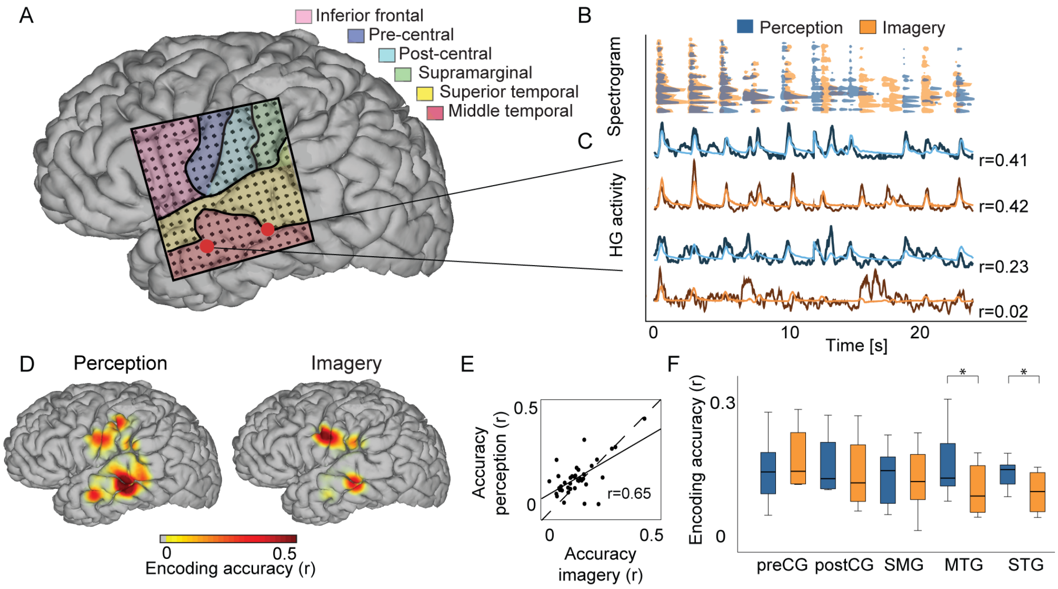


Figure 5-2. Encoding accuracy *(A) Electrode location overlaid on cortical surface reconstruction of the participant's cerebrum.* ***(B)*** *Overlay of the spectrogram contours for the perception (blue) and imagery (orange) condition (10% of maximum energy from the spectrograms).* ***(C)*** *Actual and predicted high gamma band power (70–150 Hz) induced by the music perception and imagery segment in (B). Top electrodes have very similar predictive power, whereas bottom electrodes are very different for perception and imagery. Recordings are from two different STG sites, highlighted in pink in (A).* ***(D)*** *Encoding accuracy is plotted on the cortical surface reconstruction of the participant's cerebrum (map thresholded at p<0.05; FDR correction).* ***(E)*** *Encoding accuracy of significant electrodes of the perception model as a function the imagery model. Electrode-specific encoding accuracy is correlated between both perception and imagery models (r=0.65; p<10^{-4}; randomization test).* ***(F)*** *Encoding accuracy as a function of anatomic location (pre-central gyrus (pre-CG), post-central gyrus (post-CG), supramarginal gyrus (SMG), medial temporal gyrus (MTG) and superior temporal gyrus (STG)).*

To further investigate anatomical similarities and differences between the perception and imagery conditions, we plotted the anatomical layout of prediction accuracy of individual electrodes. In both conditions, results showed that sites with the highest prediction in both conditions were located in the superior and middle temporal gyrus, pre- and post-central gyrus, and supramarginal gyrus (*Figure 5-2 D*; heat map thresholded to p<0.05; one-sample *z*-test; FDR correction), consistent with previous ECoG results (Pasley et al. 2012). Among the 256 electrodes recorded, 210 were fitted in the encoding model, while the remaining 46 electrodes were removed due to excessive noise. Within the fitted electrodes, while 35 and 15 electrodes were significant in the perception and imagery condition, respectively (p<0.05; one-sample *z*-test; FDR correction), of which nine electrodes were significant in both conditions. Anatomic locations of the electrodes with significant encoding accuracy are depicted in *Figure 5-3* and *Supplementary Figure 5-1*. To compare the encoding accuracy across conditions, we performed additional analysis on the electrodes that had significant encoding accuracy in at least one condition (41 electrodes; unless otherwise stated). Prediction accuracy of individual electrodes was correlated between perception and imagery (*Figure 5-2 E*; 41 electrodes; r=0.65; p<10^{-4}; randomization test). Because both perception and imagery models are based on the same auditory stimulus representation, the correlated prediction accuracy provides strong evidence for a shared neural representation of sound based on spectrotemporal features.

To assess how brain areas encoding auditory features varied across experimental conditions, we analyzed the significant electrodes in the gyri highlighted in *Figure 5-2 A* (pre-central gyrus (pre-CG), post-central gyrus (post-CG), supramarginal gyrus (SMG), medial temporal gyrus (MTG) and superior temporal gyrus (STG)) using Wilcoxon signed-rank test (p>0.05; one-sample Kolmogorov-Smirnov test; *Figure 5-2 F*). Results showed that the encoding accuracy in the MTG and STG was higher for the perception *(*MTG: *M* = 0.16, STG: *M = 0.13)* than for the imagery *(*MTG: *M* = 0.11, STG: *M*= 0.08; p<0.05; Wilcoxon signed-rank test; Bonferroni correction). The encoding accuracy in the pre-CG, post-CG and SMG was not different between the perception (pre-CG *M*=0.17; post-CG *M*=0.15; SMG *M*=0.12; p>0.5; Wilcoxon signed-rank test; Bonferroni correction) and imagery (pre-CG *M*=0.14; post-CG *M*=0.13; SMG *M*=0.12; p>0.5; Wilcoxon signed-rank test; Bonferroni correction) conditions. The significant improvement of the perception vs. imagery model was specific to the temporal lobe, which may reflect underlying differences in spectrotemporal encoding mechanisms, or alternatively, a greater sensitivity to discrepancies between the actual content of imagery and the recorded sound stimulus used in the model.

5.4.2 Spectrotemporal tuning during auditory perception and imagery

Auditory imagery and perception are distinct subjective experiences yet both are characterized by a sense of sound. How does the auditory system encode sound during music perception and imagery? Examples of standard STRFs are shown in *Figure 5-3* for temporal electrodes (*Supplementary Figure 5-2* for all the STRFs). These STRFs highlight neural stimulus preferences as shown by the excitatory (warm color) and inhibitory (cold color) subregions.

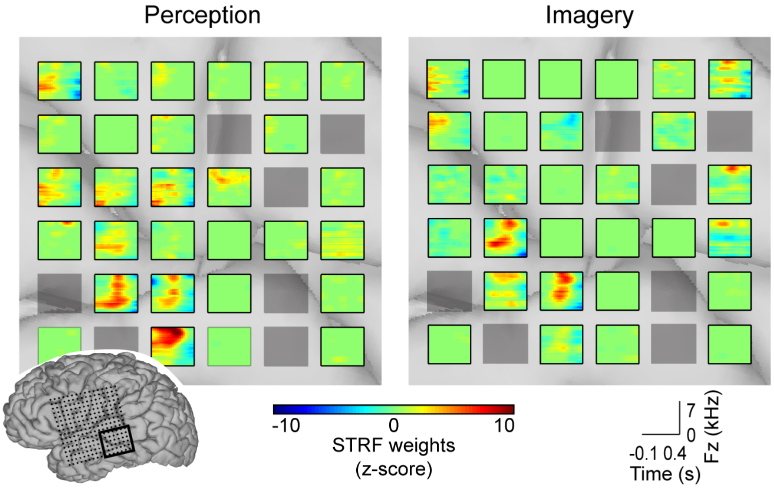


Figure 5-3 Spectrotemporal receptive fields (STRFs). *Examples of standard STRFs for the perception (left panel) and imagery (right panel) models (warm colors indicate where the neuronal ensemble is excited, cold colors indicate where the neuronal ensemble is inhibited). On the lower left corner, Electrode location overlaid on cortical surface reconstructions of the participant's cerebrum. Electrodes whose STRFs are shown are outlined in black. Grey electrodes were removed from the analysis due to excessive noise (see Materials and Methods).*

Figure 5-4 A shows the latency (s) for the perception and imagery conditions, defined as the temporal coordinates of the maximum deviation in the STRF. The peak latency was significantly correlated

between both conditions (r=0.43; p<0.005; randomization test), suggesting that both perception and imagery are simultaneously active for most of their response durations. Then, we analyzed frequency tuning curves estimated from the STRFs (see Material and methods for details). Examples of tuning curves for both perception and imagery encoding models are shown for the electrodes indicated by the black outline in the anatomic brain (*Figure 5-4 C*). Across conditions, the majority of individual electrodes exhibited a complex frequency tuning profile. For each electrode, similarities between the tuning curves in the perception and imagery models were quantified using Pearson's correlation coefficient. The anatomical distribution of tuning curve similarity is plotted in *Figure 5-4 B*, with the correlation at individual sites ranging between r=-0.3-0.6. The effect of anatomical location (pre-CG, post-CG, SMG, MTG and STG) on tuning curve similarity was not significant (*Chi-square*=3.59; p>0.1; Kruskal-Wallis Test). Similarities in tuning curve shape between auditory imagery and perception suggest a shared auditory representation, but there is no evidence that similarities depended on gyral area.

Different electrodes are sensitive to different acoustic frequencies important for music processing. We next quantitatively assessed how frequency tuning of predictive electrodes (N=41) varied during the two conditions. First, to evaluate how the acoustic spectrum was covered at the population level, we quantified the proportion of significant electrodes with a tuning peak at each acoustic frequency (*Figure 5-4 B*). Tuning peaks were identified as significant parameters in the acoustic frequency tuning curves (z>3.1; p<0.001; separated by more than one third an octave). The proportion of electrodes with tuning peaks was significantly larger for the perception (mean = 0.19) than for the imagery (mean = 0.14) condition (*Figure 5-4 C*; p<0.05; Wilcoxon signed-rank test). Despite this, both conditions exhibited reliable frequency selectivity, as nearly the full range of the acoustic frequency spectrum was encoded. The fraction of acoustic frequency covered with peaks by predictive electrodes was 0.91 for the perception and 0.89 for the imagery.

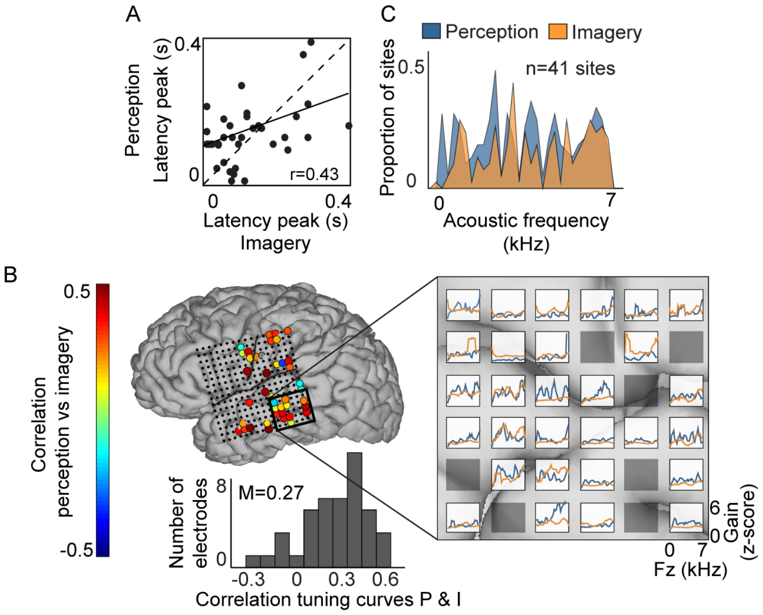


Figure 5-4 Auditory tuning. ***(A)*** *Latency peaks – estimated from the STRFs – were significantly correlated between perception and imagery conditions (r=0.43; p<0.005; randomization test).* ***(B)*** *Examples of tuning curves for both perception and imagery encoding models defined as the average gain of the STRFs as a function of acoustic frequency. Black outline in the anatomic brain indicate electrode location. for the electrodes indicated by the black outline in the left panel – Correlation coefficients between the perception*

and imagery conditions are plotted for significant electrodes on the cortical surface reconstruction of the participant's cerebrum. Grey electrodes were removed from the analysis due to excessive noise. Bottom panel is a histogram of electrode correlation coefficients between the perception and imagery tuning. ***(C)*** *Proportion of predictive electrode sites (N=41) with peak tuning at each frequency. Tuning peaks were identified as significant parameters in the acoustic frequency tuning curves (z>3.1; p<0.001) – separated by more than one third an octave.*

5.4.3 Reconstruction of auditory features during music perception and imagery

To evaluate the ability to identify piano keys from the brain activity, we reconstructed the same auditory spectrogram representation used in the encoding models. Results showed that the overall reconstruction accuracy was higher than zero in both conditions (*Figure 5-5 A*; $p < 0.001$; randomization test), but did not differ between conditions ($p > 0.05$; two-sample t-test). As a function of acoustic frequency, mean accuracy ranged from r=0–0.45 (*Figure 5-5 B*).

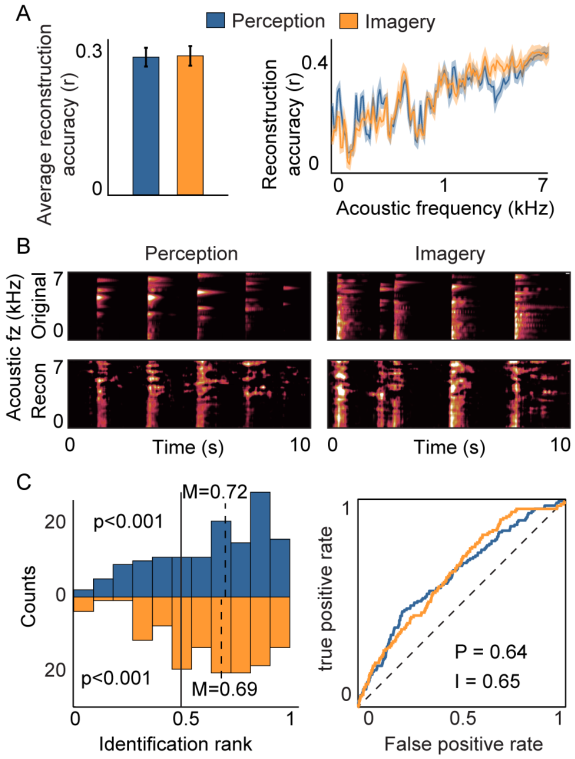


Figure 5-5 Reconstruction accuracy. ***(A)*** *Right panel: Overall reconstruction accuracy of the spectrogram representation for both perception (blue) and imagery (orange) conditions. Error bars denote resampling SEM. Left panel: Reconstruction accuracy as a function of acoustic frequency. Shaded region denotes SEM over the resamples.* ***(B)*** *Examples of original and reconstructed segments for the perception (left) and the imagery (right) model.* ***(C)*** *Distribution of identification rank for all reconstructed spectrogram notes. Median identification rank is 0.65 and 0.63 for the perception and imagery decoding model, respectively, which is significantly higher than 0.50 chance level (p<0.001; randomization test). Left panel: Receiver operating characteristic (ROC) plot of identification performance for the perception (blue curve) and imagery (orange curve) model. Diagonal black line indicates no predictive power.*

We further assessed reconstruction accuracy by evaluating the ability to identify isolated piano notes from the test set auditory spectrogram reconstructions. Examples of original and reconstructed segments are depicted in *Figure 5-5 B* for the perception (left) and imagery model (right). For the identification, we extracted 0.5-second segments at piano note onsets from the original and reconstructed auditory spectrogram. Onsets of the notes were defined with the MIRtoolbox (Lartillot, Toiviainen, and Eerola 2008). For a target segment, a similarity score (correlation coefficient) was computed between the reconstruction and the actual auditory spectrograms of each of the segments in the candidate set. The similarity scores were sorted, and identification rank was quantified as the percentile rank of the correct segment (1.0 indicates the target reconstruction matched the correct original segment out of all candidate segments; 0.0 indicates the target was least similar to the correct original segment among all other candidates). The expected mean of the distribution of identification ranks is 0.5 at chance level. Results showed that the median identification rank of individual piano notes was significantly higher than chance for both conditions (*Figure 5-5 C*; median identification rank perception = 0.72 and imagery = 0.69; p< 0.001; randomization test). Similarly, the area under the curve (AUC) of identification performance for the perception (blue curve) and imagery (orange curve) model was well above chance level (diagonal black dashed line indicates no predictive power; p<0.001; randomization test).

5.4.4 Cross-condition analysis

Another method to evaluate the overlapping degree between both perception and imagery conditions is to apply the decoding model built in the perception condition to imagery neural data, and vice-versa. This approach is based on the hypothesis that both tasks share neural mechanisms and is useful when one of the models cannot be built directly, because of the lack of observable measures. This technique has been successfully applied to various fields, such as vision (Haynes and Rees 2005; Horikawa et al. 2013; Reddy, Tsuchiya, and Serre 2010), and speech (Martin et al. 2014). When the model was trained on the perception condition and tested on the imagined condition (r=0.28), decoding performances improved by 50% compared to when the perception model was applied to imagined data (r=0.19) (*Figure 5-6*). This highlight the importance of having a model that is specific to each condition.

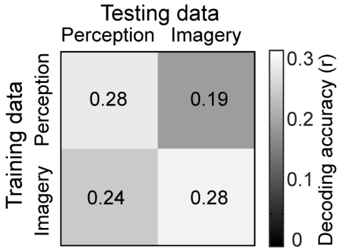


Figure 5-6 Cross-condition analysis. *Reconstruction accuracy when the decoding model was built on the perception condition and applied to the imagery neural data and vis-versa. Decoding performances improved by 50% when the model was trained on the perception condition and tested on the imagined condition (r=0.28), compared to when the perception model was applied to imagined data (r=0.19).*

5.4.5 Control analysis for sensorimotor confounds

In this study, the participant played piano in two different conditions (music perception and imagery). In addition to the auditory percept, arm-, hand- and finger-movements related to the active piano task could have presented potential confounds to the decoding process. We controlled for possible motor confounds in three different ways. First, the electrode grid did not cover hand or finger sensorimotor brain area (*Figure 5-7A*). This reduces the likelihood that the decoding model involved hand-related motor confounds. Second, examples of STRFs for two neighboring electrodes show different weight patterns for both conditions (*Figure 5-7B*). For instance, the weights of the electrode depicted in the left example of *Figure 5-7B* are correlated between both conditions (r=0.48), whereas the weights are not correlated in the right example (r=0.04). Differences across conditions cannot be explained by motor confounds, because finger movement was similar in both tasks. Third, brain areas that significantly encoded music perception and imagery overlapped with auditory sensory areas (*Supplementary Figure 5-1* and *Supplementary Figure 5-1*), as revealed by the encoding accuracy and STRFs during passive listening to TIMIT sentences (no movements). The presence of STRFs in the sensorimotor cortex is in accordance with previous research showing that the motor cortex represents acoustic features of sounds similarly to auditory cortex (Wilson et al. 2004; Cheung et al. 2016) These findings provide evidence against motor confounds, and suggest that the brain responses were induced by auditory percepts rather than motor movements associated with pressing piano keys. Finally, we built two additional decoding models, using 1) only temporal lobe electrodes and 2) only auditory-responsive electrodes (*Figure 5-7C*; see Material and methods for details). Both models showed significant reconstruction accuracy (p<0.001; randomization test) and median identification rate (p<0.001; randomization test). This suggests that even if we removed all electrodes that are potentially not related to auditory processes, the decoding model still performs above chance.

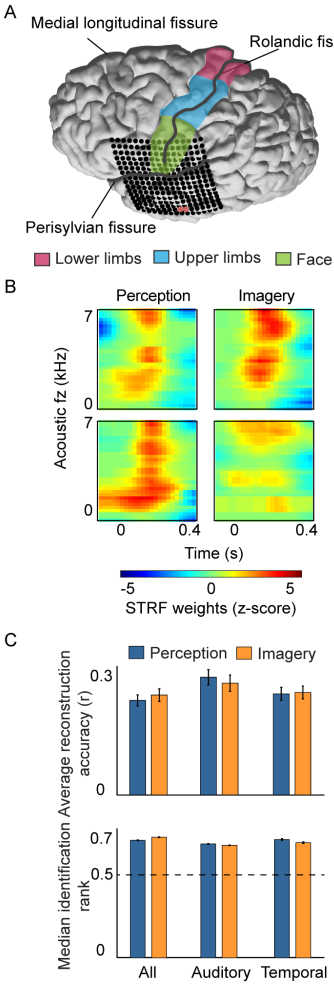


Figure 5-7 Control analysis for motor confound. ***(A)*** *Electrode location overlaid on cortical surface reconstructions of the participant's cerebrum. View from the top of the brain shows that hand motor areas not recorded with the grid. Brain areas associated with face, upper and lower limbs are depicted in green, blue and pink, respectively.* ***(B)*** *Different STRFs for two neighboring electrodes highlighted in (A) for both perception and imagery encoding models.* ***(C)*** *Overall reconstruction accuracy (left panel) and median identification rank (right panel) when using all electrodes, only temporal electrodes, and auditory electrodes (see Materials and methods for details).*

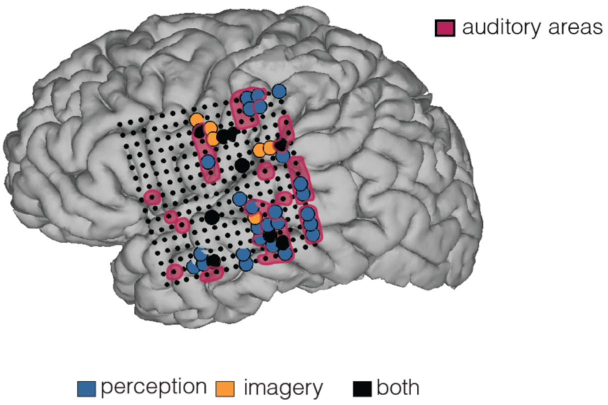
5.5 Discussion

Music imagery studies present several obstacles due to the subjective nature and absence of verifiable and observable measures. Our task design allowed precise time-locking between the recorded neural activity and spectrotemporal features of music imagery, and provided a unique opportunity to quantitatively study neural encoding during auditory imagery, and compare tuning properties with auditory perception. Here, we provide the first evidence of spectrotemporal receptive field and auditory features encoding in the brain during music imagery, providing comparative measures with actual music perception encoding. We observed that neuronal ensembles were tuned to acoustic frequencies during imagined music, suggesting that spectral organization occurs in the absence of actual perceived sound. Supporting evidence has shown that restored speech – when a speech instance is replaced by noise, but the listener perceives a specific speech sound – is grounded in acoustic representations in the superior temporal gyrus (Leonard et al. 2016). In addition, the results showed substantial, but not complete overlap in neural properties – i.e. spectral and temporal tuning properties, and brain areas – during music perception and imagery. Such findings are consistent with conclusions that visual imagery involves many, but not all, of the brain areas involved in visual perception (e.g., Kosslyn and Thompson, 2000, 2003). We also showed that auditory features could be reconstructed from neural activity of the imagined music. Because both perception and imagery models are based on the same auditory stimulus representation, the correlated prediction accuracy provides strong evidence for a shared neural representation of sound based on spectrotemporal features. This confirms that the brain encodes spectrotemporal properties of sounds – as previously shown by behavioral and brain lesion studies (see Hubbard, 2010 for a review).

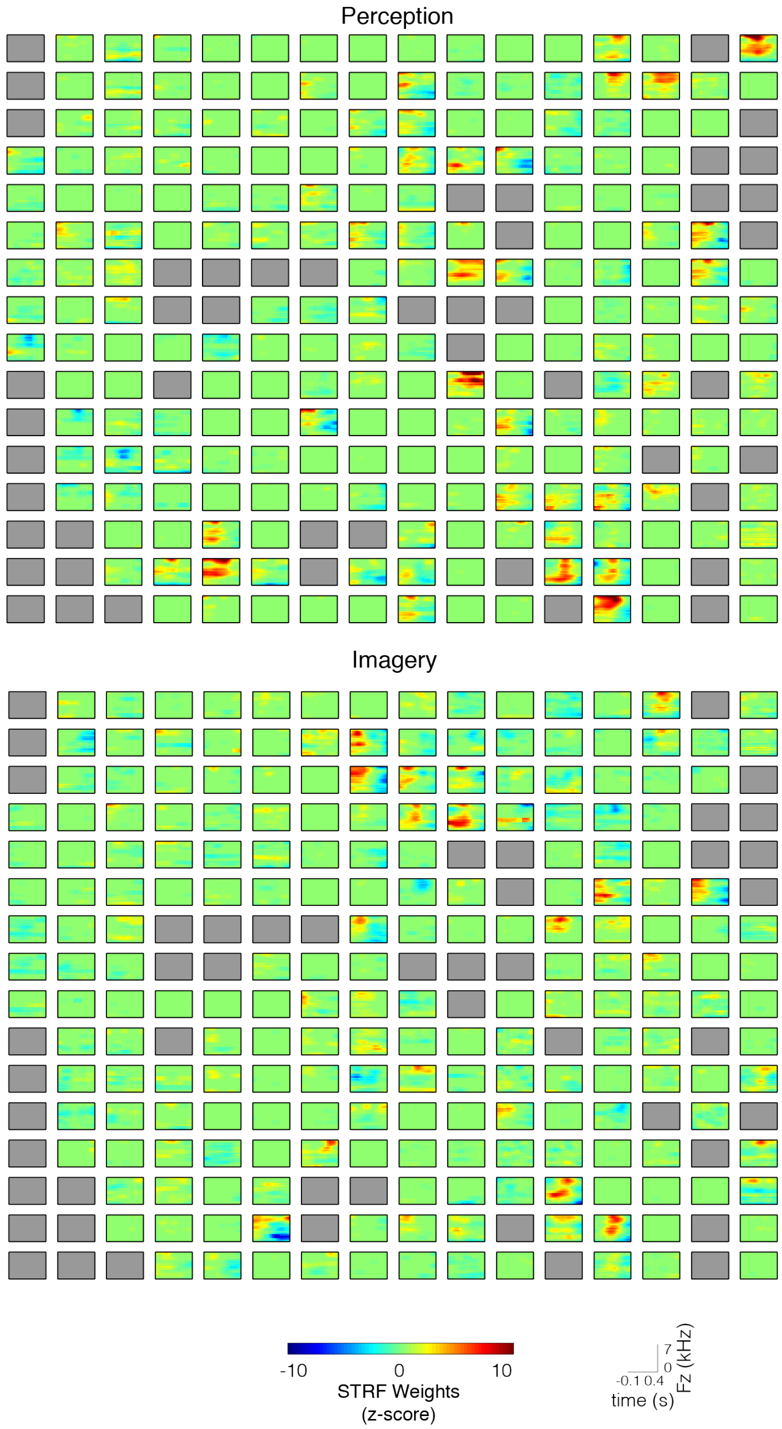
Methodological issues in investigating imagery are numerous, including the lack of evidence that the desired mental task was operational. The current task design did not allow verifying how the mental task was performed, yet the behavioral index of keynote press on the piano indicated the precise time and frequency content of the intended imagined sound. In addition, we recorded a skilled piano player, and it has been suggested that participants with musical training exhibit better pitch and temporal acuity in auditory imagery than participants with little or no musical training (Herholz et al., 2008; Janata and Paroo, 2006). Furthermore, tonotopic maps located in the STG are enlarged within trained musicians (Pantev et al. 1998). Thus, having a trained piano-player suggests improved auditory imagery ability (see also (Halpern 1988; Zatorre and Halpern 1993; Zatorre et al. 1996), and reduced issues related to spectral and temporal errors.

Finally, the electrode grid was located on the left hemisphere of the participant. This raises the question of lateralization in the brain response to music perception and imagery. Studies have shown the importance of both hemispheres for auditory perception and imagination (Griffiths 1999; Halpern et al. 2004; Kraemer et al. 2005; Rauschecker 2001; Zatorre and Halpern 1993; Zatorre et al. 1996), although brain patterns tend to shift toward the right hemisphere during music processing (see Zatorre and Halpern, 2005 for a review). In our task, the grid was located on the left hemisphere, and allowed significant encoding and decoding accuracy within high gamma frequency ranges. This is consistent with the notion that music auditory processes are also evident in the left hemisphere.

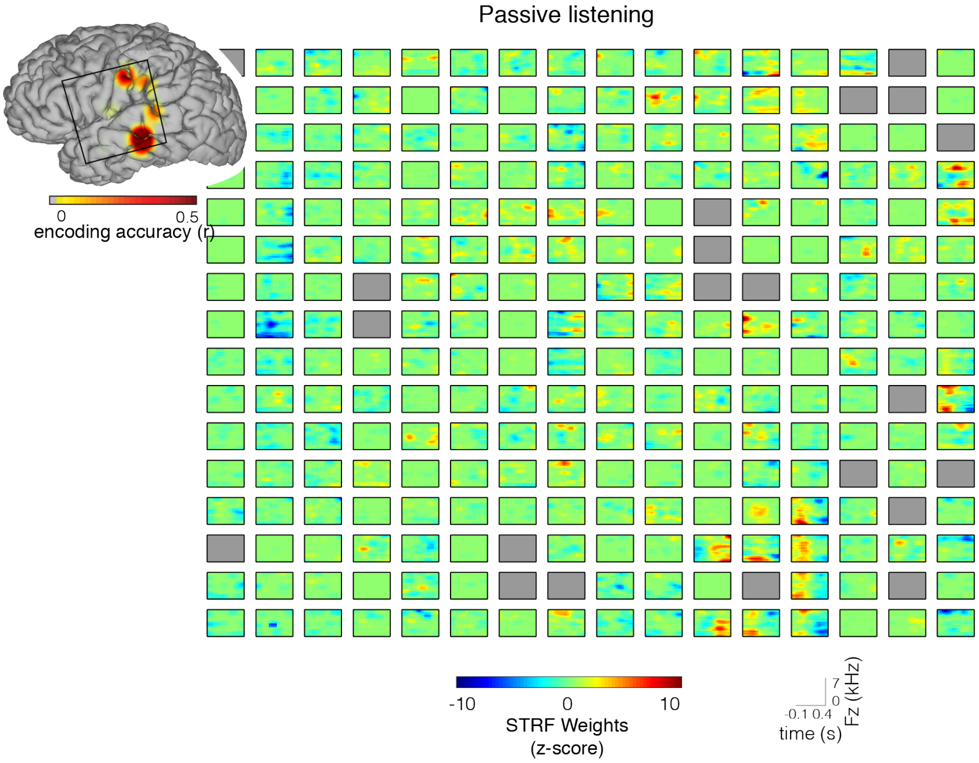
5.6 Supplementary information



Supplementary Figure 5-1 Anatomical distribution of significant electrodes. *Electrodes with significant encoding accuracy overlaid on cortical surface reconstruction of the participant's cerebrum. To define auditory sensory areas (pink), we built an encoding model on data recorded while participant listened passively to speech sentences from the TIMIT corpus (Garofolo 1993). Electrodes with significant encoding accuracy (p<0.05; FDA correction) are highlighted.*



Supplementary Figure 5-2 Spectrotemporal receptive fields. *STRFs for the perception (top panel) and imagery (bottom panel) models. Grey electrodes were removed from the analysis due to excessive noise.*



Supplementary Figure 5-3 Neural encoding during passive listening. *Standard STRFs for the passive listening model – built on data recorded while the participant listened passively to speech sentences from the TIMIT corpus (Garofolo 1993). Grey electrodes were removed from the analysis due to excessive noise. Encoding accuracy is plotted on the cortical surface reconstruction of the participant's cerebrum (map thresholded at p<0.05; FDR correction).*

Chapter 6 General discussion and conclusions

Disclaimer: This chapter is adapted from the following articles – with permissions of all co-authors and journals:

Martin S., Millán J.d.R., Knight R.T., and Pasley B.N. 2016. "The Use of Intracranial Recordings to Decode Human Language: Challenges and Opportunities." Brain and Language. doi:10.1016/j.bandl.2016.06.003.

My contribution: Conceptualization, visualization, writing – original draft preparation

Neuroengineered technologies have made tremendous advances in decoding motor or visual neural signals for assisting and restoring lost functions. However, they have failed to improve natural communication for patients with disabling neurological conditions. A few brain-computer interfaces have allowed relevant communication applications, such as moving a cursor on the screen (Wolpaw et al. 1991) and spelling letters (Farwell and Donchin 1988; Perdikis et al. 2014; Vansteensel et al. 2016; Pandarinath et al. 2017). Although this type of interface has proven to be useful, patients had to learn to modulate their brain activity in an unnatural and unintuitive way – i.e. performing mental tasks like a rotating cube, mental calculus or movements attempts to operate an interface (Millán et al. 2009) or detecting rapidly presented letter on a screen (Nijboer et al. 2008). As an alternative solution, we evaluated the feasibility to decoding directly neural correlates associated with internal speech as an input for a real-time speech prosthesis. For this, we explored various neural representations during imagined speech using electrocorticographic neural signals, and compared their relation to speech perception and production.

Throughout the discussion, we first briefly summarize the various studies undertaken and the main findings achieved. Then, we discuss how this work has contributed to new advances in the field, and opportunities to carry out innovative research. Finally, we outline the challenges faced when decoding human speech, and new avenues to push the boundaries of what is currently possible in this field of speech decoding.

6.1 Summary

This thesis was an explorative work aiming at better understanding imagined speech using electrocorticographic neural signals recorded in epileptic patients. We investigated various speech representations, such as acoustic sound features, phonemic features, and individual words. We also evaluated the ability to decode these speech features for targeting communication devices. For this, four different studies have been performed.

In **Chapter 2**, we reconstructed for the first time continuous acoustic features from high gamma neural activity recorded during imagined speech. For this, we used cross-condition linear regression, and thereby extended the mathematical framework used in Pasley et al. (2012) to imagined speech. Results showed that spectrotemporal features of imagined speech were significantly reconstructed from models built from overt speech data. This highlighted that overt speech and imagined speech share a partially common spectrotemporal neural representation in the motor cortex and perisylvian areas.

In **Chapter 3**, we decoded continuous phoneme sequences from high gamma neural activity recorded during imagined speech. In order to label intended phonemes more accurately during imagined speech, we designed a karaoke-like task, in which visual words scrolling on the screen were divided into their phonemic representations. Until now, isolated phonemes were successfully decoded during imagined speech (Ikeda et al. 2014; Pei et al. 2011; Brumberg et al. 2011), but these study failed to decode phoneme sequences during continuous speech. For this, we used hidden Markov models in order to incorporate both, a phoneme likelihood model and a language model. This approach has been widely used in the field of speech recognition (Rabiner 1993), and more recently in neural-based speech recognition (Moses et al. 2016; Herff et al. 2015). Here, we replicated these results, and extended the approach to imagined speech. Initial results in two patients were promising, nevertheless findings need to be extended to a larger pool of participants, in order to draw conclusions.

In **Chapter 4**, we classified for the first time individual words from high gamma neural time features recorded during an imagined speech word repetition task. For this, we proposed a new approach that takes time features, and deals with speech production irregularities by introducing temporal alignment in the classification framework. Although words have been decoded during overt speech (Blakely et al. 2008), only phonemes were successfully predicted during imagined speech (Ikeda et al. 2014; Pei et al. 2011; Brumberg et al. 2011). This study represents a proof of concept for basic decoding of speech imagery, yet the results to date are not yet robust enough for a clinical communication device. The major difficulties derive from the weak signal-to-noise ratio and the lack of temporal alignment across trials. For instance, in the overt speech condition, decoding performances were increased when trials were extracted at speech onset/offset compared to when trials where extracted at cue onset. Finding behavioral or neural metrics that help defining speech onset/offset in the imagined condition would improve performances.

In these studies, we investigated imagined speech in parallel with overt speech production and/or speech perception. This allowed comparing speech representations across conditions, and integrate imagined speech into the general speech network. Results revealed complex patterns of brain activity across conditions and tasks. Altogether, the most informative areas to decode imagined speech units were located in the superior temporal gyrus, inferior frontal gyrus and sensorimotor cortex, areas commonly associated with speech. However, different tasks involve different speech production processes, ranging from lexical retrieval to phonological or even phonetic encoding (Perrone-Bertolotti et al. 2014), making it difficult to draw any conclusion about the specific function of anatomic locations. In addition, the signal was significantly weaker in the imagined speech condition than in the listening or overt speech conditions, where speech stimuli were directly observed. Finally, variability across participants in the imagined condition might reflect the subjective strategy employed by each individual to generate internal speech. In sum, it is still unclear how the content of imagined speech is processed in the human cortex.

In **Chapter 5**, we investigated the neural encoding of acoustic features during music imagery. This study relied on an extremely rare clinical case in which a patient undergoing neurosurgery for epilepsy treatment was also an adept piano player. Evidence has shown that music and speech share common brain networks (Schön et al. 2010; Callan et al. 2006), and therefore helped understanding features of inner subjective experiences. While previous brain imaging studies have indicated anatomical regions active during auditory imagery (Zatorre et al. 1996; Griffiths 1999; Halpern and

Zatorre 1999; Rauschecker 2001; Halpern et al. 2004; Kraemer et al. 2005), it was unknown how fine-scale neural tuning of sound frequency were represented. This study provided a unique opportunity to apply receptive field modeling techniques to quantitatively study neural encoding during music imagery. Results showed that music perception and imagery share partial neural encoding mechanisms, a feature common to speech neural activity. Furthermore, these findings also demonstrate that receptive field and decoding models – typically applied in neuroprosthetics for motor and visual restoration – are now applicable to auditory imagery. This represents a major advance with direct application to the field of neural interfaces for restoration of communication

6.2 Opportunities

This line of inquiry demonstrated the potential of using neural predictive models as research tools to derive data driven conclusions underlying complex speech representations. In particular, we showed that encoding models have tremendous potential for uncovering the link between imagined speech representations and neural responses. Using quantitative, model-based characterizations, we showed for the first time that brain activity is tuned to various levels of speech descriptions, broken down into anatomic and functional stages.

The potential of encoding models extends beyond its established relevance in cognitive neuroscience. For instance, here, we investigated speech functions in participants with no major language deficits. However, aphasic patients have diverse language components affected, such as auditory, phonological, or lexical function, and can occur in any linguistic modality. Encoding models offer a functional explanation for specific language disorder, and allow measuring continuous changes in cortical representations (Pasley and Knight 2013). Targeted rehabilitation would benefit from quantitative measures of plasticity for guiding training-induced changes in specific cortical areas, and is applicable to a variety of aphasic symptoms having different level of speech representation affected. Similarly, the various types of language deficits exemplify the challenge in building specific speech prosthesis that addresses individual needs. In this regard, encoding models offer a unique opportunity to identify injured neural circuits.

Once damaged and healthy brain functions are identified with encoding models, decoding models can be used for the design of effective speech prostheses. In particular, we showed the feasibility to decode various speech representations during imagined speech, such as acoustic features, phonetic representations, and individual words. This suggests that various strategies and designs could be employed for building a natural communication device, depending on specific, residual speech functions. Every speech representation has pros and cons for targeting speech devices. For instance, decoding acoustic features opens door to brain-based speech synthesis, in which audible speech is synthetized directly from decoded neural patterns. This approach has already been demonstrated, where predicted speech was synthesized, and acoustically fed back to the user (Guenther et al. 2009; Brumberg et al. 2010). Yet the understandability of the produced speech sounds and the best speech parameters to model remain to be demonstrated. Alternatively, decoding units of speech, such as phonemes or words provides greater naturalness, but the optimal speech unit size to be analyzed, is still a matter of debate – e.g. the longer the unit, the larger the database needed to cover the required domain, while smaller units offer more degrees of freedom, and can build a larger set of complex utterances, as shown in (Moses et al. 2016; Herff et al. 2015). A tradeoff is the decoding of a limited vocabulary of words, which carry specific semantic information, and would be relevant in a basic clinical setting ('hungry', 'thirsty', 'yes', 'no', etc.). ¨

Although this work has revealed the potential in investigating speech with predictive models, it also emphasizes that performances currently remain insufficient to build a realistic brain-based device.

Indeed, our studies, as well as other studies on imagined speech decoding (Ikeda et al. 2014; Pei et al. 2011; Brumberg et al. 2011), reported marginal decoding accuracy, limiting their relevance for clinical scenarios. Numerous challenges were encountered, such as weak signal-to-noise-ratio, lack of behavioral output and speech irregularities, precluding robust predictions. We address these concerns and potential avenues for improvements in the section 6.3 below.

Meanwhile, an alternative to a speech-interface based solely on brain decoding is to build a hybrid system, which acquires sensor data from multiple elements of the human speech production system, and combine the different signals to optimize speech synthesis (see Brumberg et al. 2010 for a review). For instance, recording sensors allow characterizing the vocal tract by measuring its configuration directly or by sounding it acoustically using electromagnetic articulography, ultrasound or optical imaging of the tongue and lip. Alternatively, electrical measurements can infer articulation from actuator muscle signals (i.e., using surface electromyography) or signals obtained directly from the brain (mainly EEG and ECoG). Using different sensors and different speech representations allow exploiting an individual's residual speech functions to operate the speech synthesis.

Unique opportunities for targeting communication assistive technologies are offered by combining different research fields. Neuroscience reveals which anatomical locations and brain signals should be modeled. Linguistic fields support development of decoding models that incorporate linguistic, contextual specifications – including segmental elements and supra-segmental elements. Combining insights from these research fields with machine learning and speech recognition algorithms is a key element to improve prediction accuracies. Finally, the success of speech neuroprosthesis depends on the continuous technological improvements to enhance signal quality and resolution, and allow developing more portable and biocompatible invasive recording devices. Merging various fields together will allow tackling the challenges central to decoding imagined speech.

6.3 Challenges and solutions

In this section, we highlight numerous challenges that were encountered, such as dealing with internal neural representations and facing technological limitations. For each subsection, we first define what was the challenge, then we describe how we tried to tackle it, and finally, we provide possible improvements that can be made.

Dealing with internal neural representations

Physiological studies have unraveled many neural mechanisms of speech perception and production that occur at a very fine temporal scale, such as acoustic processing in the early auditory periphery, phonetic encoding in posterior areas of the temporal lobe and semantic and higher level of linguistic processes in the anterior areas of the temporal lobe (Chang et al. 2010; Mesgarani et al. 2014; Bouchard et al. 2013, 2013). Conversely, the limited physiological studies on imagined speech emphasizes the difficulties when investigating speech representations during internal subjective experiences like imagined speech.

Reasons for this are numerous, and include the lack of behavioral output and impossibility of monitoring precisely the fine spectrotemporal fluctuation property of speech. Critically, imagined speech cannot be directly observed by an experimenter. As a consequence, it is complicated to time-lock brain activity to a measurable stimulus or behavioral state, and to build predictive models that directly regress the neural activity to any behavioral metric or speech representation. In addition, natural speech expression is not just operated under conscious control, but is affected by various factors, including gender, emotional state, tempo, pronunciation and dialect, resulting in temporal irregularities (stretching/compressing, onset/offset delays) across repetitions. As a result, this leads to problems in exploiting the temporal resolution of electrocorticography to investigate imagined speech.

We tried alleviating these problems by designing tasks that maximize the accuracy when labeling the content of imagery. For instance, we tried to cue the participants in a rhythmical manner, and designed a karaoke-like task that allowed participants to prepare themselves, and be more consistent. Despite these task design efforts, results showed inconsistencies between the actual cue and the speech onset/offset in the overt condition.

A key issue related to this is that patients were not familiarized with our tasks before entering in the epilepsy monitoring unit. Studies have shown that participants with musical training exhibited better pitch and temporal acuity in auditory imagery and enlarged tonotopic maps located in the STG than did participants with little or no musical training (Herholz et al., 2008; Janata and Paroo, 2006; Pantev et al. 1998). As such, we argue it would be beneficial to train subjects on speech imagery, in order for them to be more consistent in time and way of performing the imagined speech. A possible training scenario is to incorporate an online feedback in the protocol. As such, closing the loop could enhance the task performances, reinforce the neural signal and maximize consistency.

In addition, finding a behavioral or neural metric that allows marking more precisely the imagined speech onset and offset would reduce the temporal uncertainty window and measure imagined performances. This will be increasingly important when we move towards asynchronous protocols, i.e. when patients spontaneously produce imagined speech, as opposed to current protocols in which they are cued. To this end, a potential biomarker is the high gamma neural activity, which demonstrated the ability to define an active speech window in *Chapter 3*. Other representative examples come from behavioral and psychology studies, which have relied on indirect measures to infer the existence and properties of the intended imagined experience. Various examples of behavioral measures have provided convincing evidence that internal imagery was actually generated, e.g. subjective report of participants or comparisons of performances on task selective facilitation (see Hubbard 2010 for a complete review). Therefore, monitoring performances and vividness during imagined speech might alleviate pronunciation and grammatical mistakes, while maximizing the signal-to-noise ratio.

6.3.1 Elucidating the neurobiology of language

Understanding speech processing is a key step to building efficient natural speech prosthesis, yet the complex neural mechanisms underlying speech remain largely unknown. A reason for this is that speech is exclusive to humans and cannot be studied in animal models, in comparison to other extensively studied cognitive domains, such as perception, memory and decision making. Animals use a system of communication that is believed to be limited to expression of a finite number of genetically determined utterances (Tomasello 2008). In contrast, humans can produce a large range of utterances from a finite set of elements (Trask 1999). Despite this, lower-level auditory and motor processing has been widely explored in nonhuman mammals (Georgopoulos, Kettner, and Schwartz 1988; deCharms 1998; Depireux et al. 2001) and avians (de Boer 1967; Theunissen et al. 2001). Similarly, electrophysiological evidence has shown that macaques have two separate cortical pathways (ventral and dorsal streams) for audition (Romanski et al. 1999), resembling the human speech architecture. These findings emphasize how critical bidirectional interactions between animal and human studies are for understanding the human brain (Badre, Frank, and Moore 2015).

In addition to experimental barrier, speech is organized in a widely distributed and complex network that works along different time scales. In this work, we focused on high-gamma frequency bands, as they have been shown to provide the most reliable index of local cortical activity, and widely used in speech decoding (Pasley et al. 2012; Mesgarani and Chang 2012; Houde and Chang 2015; Moses et al. 2016). However, other frequency ranges have been shown to carry essential information about various neurolinguistics features. For instance, the theta band (4-7 Hz) tracks the acoustic envelope of speech, correlates with syllabic rate, and discriminates spoken sentences (Luo and Poeppel 2007; Giraud and Poeppel 2012; Ding and Simon 2012; Zion Golumbic et al. 2013). Alternatively, the delta range (1-2 Hz) typically transform the signal input into lexical and phrasal units. Given the low performances in the various tasks investigated here, other frequency bands should be explored for speech decoding, and may provide complementary information to high frequency activity.

Another difficulty when targeting speech devices is that neural activity associated with speech is not invariable, but modulated by top-down influences based on expectations (Leonard et al. 2015; Leonard et al. 2016; Holdgraf et al. 2016), feedback monitoring loops (Chang et al. 2013; Houde and Chang 2015) attentional resources (Fritz et al. 2007; Mesgarani and Chang 2012) and lexical retrieval (Cibelli et al. 2015). For instance, linguistic context refers to the factors that affect the acoustic realization of speech sounds, including segmental elements, such as co-articulatory features, and supra-segmental elements, such as stress, prosodic patterns, phonation type, and intonation. While context-dependent modeling is common in speech recognition (Waibel and Lee 1990) and known to significantly improve recognition performances, it has rarely been taken into account for neural decoding. One reason for this is that it remains largely unknown how the brain encodes the various factors affecting the production of speech sound. A key aspect for improving speech prosthetic will be to determine which factors significantly improve decoding performances, and how to model them.

Finally, language is involved in a variety of modalities including writing, reading, listening and speaking. These four modalities share receptive and expressive areas of the brain, yet, they also have unique processing levels and neuroanatomical substrates (Berninger and Abbott 2010; Singleton and Shulman 2014). For instance, a person with a writing deficit may not be as impaired in speaking, and vice versa (Rapp, Fischer-Baum, and Miozzo 2015). Results have been highly variables, because different tasks involve different speech processes, engaging different brain modality. These observations exhibit the complexities of linguistic organization found in the brain (Bellugi, Poizner, and Klima 1989). We suggest that a complete characterization of the language network, at all the scales and including all forms of expressive and receptive mechanisms, will help shaping optimal communication devices.

6.3.2 Integrating recent progress in machine learning

A major challenge in neural decoding applications lies in the computational models used for decoding the neural features. Initial studies on speech decoding used linear decoding models to map the neural activity to the speech representations (Pasley et al. 2012; Mesgarani and Chang 2012; Mesgarani et al. 2014). However, in reality, the neural correlates of language are likely non-linearly related to the various speech representations.

Recently, electrophysiological studies on speech decoding have shown promising results by integrating knowledge from the field of speech recognition (Moses et al. 2016; Herff et al. 2015). Speech recognition has been concerned with the statistical modeling of natural language for many decades, and has faced many problems that are similar to decoding neural pattern associated with speech. As such, we argue that integrating their knowledge into our field is a necessary element to succeed in the ultimate goal of a clinically reliable speech prosthesis. For instance, speech recognition has developed methodologies that enables the recognition and translation of spoken language into text. This was achieved by incorporating extensive knowledge about how speech is produced and perceived at various phonetic levels (acoustic, auditory, articulatory features), and from advances in computer resources and big data management to build remarkable applications, such as spellcheck tool, natural speech synthesizer and translation program. Similarly, advanced probabilistic models might be more adapted in order to deal with problems associated with speech production temporal irregularities, than a deterministic approach like dynamic time warping, which is not robust for noisy data.

More complex models with increasing number of parameters can be used, but require more data to train and evaluate the models. When using electrocorticographic recordings, available data are limited. Experimental paradigms usually do not last long to avoid overloading the patients. As an alternative to traditional protocols, researchers are slowly moving toward continuous brain monitoring during the electrode implantation time. This allows increasing the amount of recorded data and is more pleasant for the participant as he is recorded in his natural, hospital environment, e.g. watch television, have interaction with relatives and clinicians, read, etc. The major problem with this approach is how to label precisely the recordings. Indeed, while it is currently possible to monitor conversations with a microphone, the continuous labeling of categories or events during a movie or a dialogue is a tedious process, and often requires a human intervention. In addition, as mentioned earlier, monitoring and labeling internal mental states, such as mood, emotions, internal speech, is problematical. We suggest that unsupervised methods might be adapted in this context, and alleviate issues associated with speech segmentation. Similarly, transfer learning, which consists in transferring the knowledge between known speech parameters.

6.3.3 Facing technological limitations

Recordings in humans are generally restricted to noninvasive techniques, such as EEG, MEG or fMRI. These approaches give large-scale overviews of cortical activity in distributed language networks, but typically lack either spatial or temporal resolution. A few intracortical recordings have shown promising results in decoding intended phonemes (Brumberg et al. 2011) and formant frequencies (Guenther et al. 2009). Although intracortical recordings have higher spatial resolution than intracranial recordings, their spatial coverage is limited, making them less suitable to investigate higher speech processing levels. Given its unique spatiotemporal properties, electrocorticography is a promising technique to decode speech, but its opportunities are limited in humans, and dependent on patients with epilepsy undergoing neurosurgical procedure for brain ablation with implanted ECoG grids. In addition, although electrocorticography provide the opportunity to investigate speech, the configuration, location and duration of implantation are not designed for the experiments, but solely for clinical purposes. In order to deal with this, we recorded the various tasks only when the grids coverage was optimal, such as electrodes were lying on the temporal cortex, the inferior temporal cortex and the sensorimotor cortex.

In addition to these clinical limitations, the design of the intracranial recording electrodes has been shown to be an important factor in motor decoding performance. Namely, the spatial resolution of a cortical surface electrode array depends on the size and spacing of the electrodes, as well as the volume of tissue to which each electrode is sensitive (Wodlinger et al. 2011). Many researchers have attempted to define what the optimal electrode spacing and size could be (Slutzky et al. 2010), but this is still an open area of research. Emerging evidence showed that decoding performances were improved when neural activity was derived from very high-density grids (Blakely et al. 2008; Rouse et al. 2013). However, although a smaller inter-electrodes spacing increases the spatial resolution, it poses additional technical issues related to the electrode grid design. We are currently investigating into high density grids with an inter-electrode spacing of 3mm. Higher density grids placed at specific speech locations would provide higher spatial resolution and potentially enhanced the signal's discriminability. Other researchers are working on increasing the number of recording contacts, having biocompatible materials and wireless telemetry for transmission of recordings from multiple electrode implants (Brumberg et al. 2011; Khodagholy et al. 2014).

Finally, long-term implantation abilities in human is lacking, as compared to non-human primate studies that showed stable neural decoding for extended periods of time (weeks to months; Ashmore et al. 2012). Reasons for these technical difficulties are the increased impedance leading to loss of signal and increase in the foreign body response to electrodes (Groothuis et al. 2014). Indeed, device material and electrode-architecture influences the tissue reaction. Softer neural implants with shape and elasticity of dura mater increase electrode conductivity and improve the implant-tissue integration (Minev et al. 2015).

To conclude, we demonstrated the potential of using predictive models to unravel neural mechanisms associated with complex cognitive functions. We also showed that various speech representations, such as acoustic features, phonemic features and individual words could be decoded from high gamma brain signals. We achieved this by designing new protocols, that ought to maximize consistencies across repetitions of the same speech instance. In addition, we introduced non-linear temporal alignments in the decoding framework, in order to deal with speech irregularities. Finally, we used state-of-the-art modeling techniques. Although, these results revealed a promising avenue for direct decoding of natural speech, they also emphasized that performance was currently insufficient to build a realistic brain-based device. We highlighted numerous challenges that likely precluded better performances, such as the low signal-to-noise-ratio, and the difficulty in monitoring precisely imagined speech. As such challenges are solved, decoding speech directly from neural activity opens the door to new communication interfaces that may allow for more natural speech-like communication in patients with severe communication deficits. We suggested new avenues that will hopefully help building a neural-based speech interface.

Chapter 7 References

Aertsen, A. M. H. J., J. H. J. Olders, and P. I. M. Johannesma. 1981. “Spectro-Temporal Receptive Fields of Auditory Neurons in the Grassfrog: III. Analysis of the Stimulus-Event Relation for Natural Stimuli.” *Biological Cybernetics* 39 (3): 195–209. doi:10.1007/BF00342772.

Alderson-Day, Ben, and Charles Fernyhough. 2015. “Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology.” *Psychological Bulletin* 141 (5): 931–65. doi:10.1037/bul0000021.

Aleman, A. 2004. “The Functional Neuroanatomy of Metrical Stress Evaluation of Perceived and Imagined Spoken Words.” *Cerebral Cortex* 15 (2): 221–28. doi:10.1093/cercor/bhh124.

Ashmore, R. C., B. M. Endler, I. Smalianchuk, A. D. Degenhart, N. G. Hatsopoulos, E. C. Tyler-Kabara, A. P. Batista, and W. Wang. 2012. “Stable Online Control of an Electrocorticographic Brain-Computer Interface Using a Static Decoder.” In , 1740–44. IEEE. doi:10.1109/EMBC.2012.6346285.

Aziz-Zadeh, Lisa, Luigi Cattaneo, Magali Rochat, and Giacomo Rizzolatti. 2005. “Covert Speech Arrest Induced by rTMS over Both Motor and Nonmotor Left Hemisphere Frontal Sites.” *Journal of Cognitive Neuroscience* 17 (6): 928–38. doi:10.1162/0898929054021157.

Badre, David, Michael J. Frank, and Christopher I. Moore. 2015. “Interactionist Neuroscience.” *Neuron* 88 (5): 855–60. doi:10.1016/j.neuron.2015.10.021.

Ball, Tonio, Markus Kern, Isabella Mutschler, Ad Aertsen, and Andreas Schulze-Bonhage. 2009. “Signal Quality of Simultaneously Recorded Invasive and Non-Invasive EEG.” *NeuroImage* 46 (3): 708–16. doi:10.1016/j.neuroimage.2009.02.028.

Basho, Surina, Erica D Palmer, Miguel A Rubio, Beverly Wulfeck, and Ralph-Axel Müller. 2007. “Effects of Generation Mode in fMRI Adaptations of Semantic Fluency: Paced Production and Overt Speech.” *Neuropsychologia* 45 (8): 1697–1706. doi:10.1016/j.neuropsychologia.2007.01.007.

Bellugi, Ursula, Howard Poizner, and Edward S. Klima. 1989. “Language, Modality and the Brain.” *Trends in Neurosciences* 12 (10): 380–88. doi:10.1016/0166-2236(89)90076-3.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. doi:10.2307/2346101.

Berninger, Virginia W., and Robert D. Abbott. 2010. “Listening Comprehension, Oral Expression, Reading Comprehension, and Written Expression: Related yet Unique Language Systems in Grades 1, 3, 5, and 7.” *Journal of Educational Psychology* 102 (3): 635–51. doi:10.1037/a0019319.

Birk, Diedenhofen. 2013. “Cocor: Comparing Correlations.” http://r.birkdiedenhofen.de/pckg/cocor/.

Blakely, Timothy, Kai J Miller, Rajesh P N Rao, Mark D Holmes, and Jeffrey G Ojemann. 2008. “Localization and Classification of Phonemes Using High Spatial Resolution Electrocorticography (ECoG) Grids.” *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 2008: 4964–67. doi:10.1109/IEMBS.2008.4650328.

Boer, E. de. 1967. “Correlation Studies Applied to the Frequency Resolution of the Cochlea.” *Journal of Auditory Research 7*.

Boonstra, T. W., S. Houweling, and M. Muskulus. 2009. “Does Asynchronous Neuronal Activity Average out on a Macroscopic Scale?” *Journal of Neuroscience* 29 (28): 8871–74. doi:10.1523/JNEUROSCI.2020-09.2009.

Bouchard, Kristofer E., Nima Mesgarani, Keith Johnson, and Edward F. Chang. 2013. “Functional Organization of Human Sensorimotor Cortex for Speech Articulation.” *Nature* 495 (7441): 327–32. doi:10.1038/nature11911.

Brodsky, Warren, Avishai Henik, Bat-Sheva Rubinstein, and Moshe Zorman. 2003. “Auditory Imagery from Musical Notation in Expert Musicians.” *Perception & Psychophysics* 65 (4): 602–12.

Brown, Steven, Angela R Laird, Peter Q Pfordresher, Sarah M Thelen, Peter Turkeltaub, and Mario Liotti. 2009. “The Somatotopy of Speech: Phonation and Articulation in the Human Motor Cortex.” *Brain and Cognition* 70 (1): 31–41. doi:10.1016/j.bandc.2008.12.006.

Brumberg, Jonathan S., Dean J. Krusienski, Shreya Chakrabarti, Aysegul Gunduz, Peter Brunner, Anthony L. Ritaccio, and Gerwin Schalk. 2016. “Spatio-Temporal Progression of Cortical Activity Related to Continuous Overt and Covert Speech Production in a Reading Task.” Edited by Stefan Elmer. *PLOS ONE* 11 (11): e0166872. doi:10.1371/journal.pone.0166872.

Brumberg, Jonathan S., Alfonso Nieto-Castanon, Philip R. Kennedy, and Frank H. Guenther. 2010. “Brain–computer Interfaces for Speech Communication.” *Speech Communication* 52 (4): 367–79. doi:10.1016/j.specom.2010.01.001.

Brumberg, Wright, Andreasen, Guenther, and Kennedy. 2011. “Classification of Intended Phoneme Production from Chronic Intracortical Microelectrode Recordings in Speech-Motor Cortex.” *Frontiers in Neuroscience*, May. doi:10.3389/fnins.2011.00065.

Bunzeck, Nico, Torsten Wuestenberg, Kai Lutz, Hans-Jochen Heinze, and Lutz Jancke. 2005. “Scanning Silence: Mental Imagery of Complex Sounds.” *NeuroImage* 26 (4): 1119–27. doi:10.1016/j.neuroimage.2005.03.013.

Buzsáki, György, Costas A. Anastassiou, and Christof Koch. 2012. “The Origin of Extracellular Fields and Currents — EEG, ECoG, LFP and Spikes.” *Nature Reviews Neuroscience* 13 (6): 407–20. doi:10.1038/nrn3241.

Callan, Daniel E., Vassiliy Tsytsarev, Takashi Hanakawa, Akiko M. Callan, Maya Katsuhara, Hidenao Fukuyama, and Robert Turner. 2006. "Song and Speech: Brain Regions Involved with Perception and Covert Production." *NeuroImage* 31 (3): 1327–42. doi:10.1016/j.neuroimage.2006.01.036.

Chakrabarti, S., Dean J Krusienski, Gerwin Schalk, and Jonathan S Brumberg. 2013. "Predicting Mel-Frequency Cepstral Coefficients from Electrocorticographic Signals during Continuous Speech Production." *6th International IEEE/EMBS Conference on Neural Engineering (NER)*. http://neuro.embs.org/files/2013/0607_FI.pdf.

Chan, A. M., A. R. Dykstra, V. Jayaram, M. K. Leonard, K. E. Travis, B. Gygi, J. M. Baker, et al. 2014. "Speech-Specific Tuning of Neurons in Human Superior Temporal Gyrus." *Cerebral Cortex* 24 (10): 2679–93. doi:10.1093/cercor/bht127.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology* 2 (3): 1–27. doi:10.1145/1961189.1961199.

Chang, E. F., C. A. Niziolek, R. T. Knight, S. S. Nagarajan, and J. F. Houde. 2013. "Human Cortical Sensorimotor Network Underlying Feedback Control of Vocal Pitch." *Proceedings of the National Academy of Sciences* 110 (7): 2653–58. doi:10.1073/pnas.1216827110.

Chang, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. 2010. "Categorical Speech Representation in Human Superior Temporal Gyrus." *Nature Neuroscience* 13 (11): 1428–32. doi:10.1038/nn.2641.

Cheung, Connie, Liberty S. Hamiton, Keith Johnson, and Edward F. Chang. 2016. "The Auditory Representation of Speech Sounds in Human Motor Cortex." *eLife* 5 (March). doi:10.7554/eLife.12577.

Chi, T, Y Gao, M C Guyton, P Ru, and S Shamma. 1999. "Spectro-Temporal Modulation Transfer Functions and Speech Intelligibility." *The Journal of the Acoustical Society of America* 106 (5): 2719–32.

Chi, T., P. Ru, and S. A. Shamma. 2005. "Multiresolution Spectrotemporal Analysis of Complex Sounds." *The Journal of the Acoustical Society of America* 118 (2): 887. doi:10.1121/1.1945807.

Cibelli, Emily S., Matthew K. Leonard, Keith Johnson, and Edward F. Chang. 2015. "The Influence of Lexical Statistics on Temporal Lobe Cortical Dynamics during Spoken Word Listening." *Brain and Language* 147 (August): 66–75. doi:10.1016/j.bandl.2015.05.005.

Clopton, B. M., and P. M. Backoff. 1991. "Spectrotemporal Receptive Fields of Neurons in Cochlear Nucleus of Guinea Pig." *Hearing Research* 52 (2): 329–44.

Corley, Martin, Paul H. Brocklehurst, and H. Susannah Moat. 2011. "Error Biases in Inner and Overt Speech: Evidence from Tongue Twisters." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37 (1): 162–75. doi:10.1037/a0021321.

Crone, N E, D Boatman, B Gordon, and L Hao. 2001. "Induced Electrocorticographic Gamma Activity during Auditory Perception. Brazier Award-Winning Article, 2001." *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 112 (4): 565–82.

deCharms, R. C. 1998. "Optimizing Sound Features for Cortical Neurons." *Science* 280 (5368): 1439–44. doi:10.1126/science.280.5368.1439.

Démonet, Jean-François, Guillaume Thierry, and Dominique Cardebat. 2005. "Renewal of the Neurophysiology of Language: Functional Neuroimaging." *Physiological Reviews* 85 (1): 49–95. doi:10.1152/physrev.00049.2003.

Depireux, D. A., J. Z. Simon, D. J. Klein, and S. A. Shamma. 2001. "Spectro-Temporal Response Field Characterization with Dynamic Ripples in Ferret Primary Auditory Cortex." *Journal of Neurophysiology* 85 (3): 1220–34.

Ding, N., and J. Z. Simon. 2012. "Emergence of Neural Encoding of Auditory Objects While Listening to Competing Speakers." *Proceedings of the National Academy of Sciences* 109 (29): 11854–59. doi:10.1073/pnas.1205381109.

Efron, Bradley. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics. http://epubs.siam.org/doi/book/10.1137/1.9781611970319.

Eggermont, J.J., A.M.H.J. Aertsen, and P.I.M. Johannesma. 1983. "Quantitative Characterisation Procedure for Auditory Neurons Based on the Spectro-Temporal Receptive Field." *Hearing Research* 10 (2): 167–90. doi:10.1016/0378-5955(83)90052-7.

Elliott, Taffeta M, and Frédéric E Theunissen. 2009. "The Modulation Transfer Function for Speech Intelligibility." *PLoS Computational Biology* 5 (3): e1000302. doi:10.1371/journal.pcbi.1000302.

Ellis, D. 2003. *Dynamic Time Warping (DTW) in Matlab*. http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/.

Farwell, L. A., and E. Donchin. 1988. "Talking off the Top of Your Head: Toward a Mental Prosthesis Utilizing Event-Related Brain Potentials." *Electroencephalography and Clinical Neurophysiology* 70 (6): 510–23.

Feinberg, T E, L J Gonzalez Rothi, and K M Heilman. 1986. "'Inner Speech' in Conduction Aphasia." *Archives of Neurology* 43 (6): 591–93.

Felton, Elizabeth A., J. Adam Wilson, Justin C. Williams, and P. Charles Garell. 2007. "Electrocorticographically Controlled Brain-Computer Interfaces Using Motor and Sensory Imagery in Patients with Temporary Subdural Electrode Implants. Report of Four Cases." *Journal of Neurosurgery* 106 (3): 495–500. doi:10.3171/jns.2007.106.3.495.

Fisher, R. A. 1915. "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population." *Biometrika* 10 (4): 507. doi:10.2307/2331838.

———. 1936. "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS." *Annals of Eugenics* 7 (2): 179–88. doi:10.1111/j.1469-1809.1936.tb02137.x.

Flinker, A., E. F. Chang, N. M. Barbaro, M. S. Berger, and R. T. Knight. 2011. "Sub-Centimeter Language Organization in the Human Temporal Lobe." *Brain and Language* 117 (3): 103–9. doi:10.1016/j.bandl.2010.09.009.

Flinker, Korzeniewska, Avgusta Y. Shestyuk, Piotr J. Franaszczuk, Nina F. Dronkers, Robert T. Knight, and Nathan E. Crone. 2015. "Redefining the Role of Broca's Area in Speech." *Proceedings of the National Academy of Sciences* 112 (9): 2871–75. doi:10.1073/pnas.1414491112.

Forney, G.D. 1973. "The Viterbi Algorithm." *Proceedings of the IEEE* 61 (3): 268–78. doi:10.1109/PROC.1973.9030.

Fritz, Jonathan B, Mounya Elhilali, Stephen V David, and Shihab A Shamma. 2007. "Auditory Attention—focusing the Searchlight on Sound." *Current Opinion in Neurobiology* 17 (4): 437–55. doi:10.1016/j.conb.2007.07.011.

Garofolo, John S. 1993. "TIMIT Acoustic-Phonetic Continuous Speech Corpus."

Georgopoulos, A P, R E Kettner, and A B Schwartz. 1988. “Primate Motor Cortex and Free Arm Movements to Visual Targets in Three-Dimensional Space. II. Coding of the Direction of Movement by a Neuronal Population.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 8 (8): 2928–37.

Geva, Correia, and Warburton. 2011. “Diffusion Tensor Imaging in the Study of Language and Aphasia.” *Aphasiology* 25 (5): 543–58. doi:10.1080/02687038.2010.534803.

Geva, Jones, Crinion, Baron, and E Warburton. 2011. “The Neural Correlates of Inner Speech Defined by Voxel-Based Lesion-Symptom Mapping.” *Brain* 134 (10): 3071–82. doi:10.1093/brain/awr232.

Geva, S., S. Bennett, E Warburton, and K Patterson. 2011. “Discrepancy between Inner and Overt Speech: Implications for Post-Stroke Aphasia and Normal Language Processing.” *Aphasiology* 25 (3): 323–43. doi:10.1080/02687038.2010.511236.

Giorgino, Toni. 2009. “Computing and Visualizing Dynamic Time Warping Alignments in R: The Dtw Package.” *Journal of Statistical Software* 31. http://www.jstatsoft.org/v31/i07.

Giraud, Anne-Lise, and David Poeppel. 2012. “Cortical Oscillations and Speech Processing: Emerging Computational Principles and Operations.” *Nature Neuroscience* 15 (4): 511–17. doi:10.1038/nn.3063.

Gönen, Mehmet, and Alpaydin Ethem. 2011. “Multiple Kernel Learning Algorithms.” *Journal of Machine Learning Research*, 12 edition.

Gracco, V. L., and A. Löfqvist. 1994. “Speech Motor Coordination and Control: Evidence from Lip, Jaw, and Laryngeal Movements.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 14 (11 Pt 1): 6585–97.

Griffiths, T. D. 1999. “Human Complex Sound Analysis.” *Clinical Science (London, England: 1979)* 96 (3): 231–34.

Groothuis, Jitte, Nick F. Ramsey, Geert M.J. Ramakers, and Geoffrey van der Plasse. 2014. “Physiological Challenges for Intracortical Electrodes.” *Brain Stimulation* 7 (1): 1–6. doi:10.1016/j.brs.2013.07.001.

Guenther, Frank H., Jonathan S. Brumberg, E. Joseph Wright, Alfonso Nieto-Castanon, Jason A. Tourville, Mikhail Panko, Robert Law, et al. 2009. “A Wireless Brain-Machine Interface for Real-Time Speech Synthesis.” Edited by Eshel Ben-Jacob. *PLoS ONE* 4 (12): e8218. doi:10.1371/journal.pone.0008218.

Gunduz, Aysegul, Peter Brunner, Amy Daitch, Eric C Leuthardt, Anthony L Ritaccio, Bijan Pesaran, and Gerwin Schalk. 2012. “Decoding Covert Spatial Attention Using Electrocorticographic (ECoG) Signals in Humans.” *NeuroImage* 60 (4): 2285–93. doi:10.1016/j.neuroimage.2012.02.017.

Halgren, Eric, Ksenija Marinkovic, and Patrick Chauvel. 1998. “Generators of the Late Cognitive Potentials in Auditory and Visual Oddball Tasks.” *Electroencephalography and Clinical Neurophysiology* 106 (2): 156–64. doi:10.1016/S0013-4694(97)00119-3.

Halpern. 1988. “Mental Scanning in Auditory Imagery for Songs.” *Journal of Experimental Psychology. Learning, Memory, and Cognition* 14 (3): 434–43.

Halpern, A. R. 1989a. “Memory for the Absolute Pitch of Familiar Songs.” *Memory & Cognition* 17 (5): 572–81.

Halpern, A R. 1989b. “Memory for the Absolute Pitch of Familiar Songs.” *Memory & Cognition* 17 (5): 572–81.

Halpern, A. R. 1999. “When That Tune Runs Through Your Head: A PET Investigation of Auditory Imagery for Familiar Melodies.” *Cerebral Cortex* 9 (7): 697–704. doi:10.1093/cercor/9.7.697.

———. 2001. “Cerebral Substrates of Musical Imagery.” *Annals of the New York Academy of Sciences* 930 (June): 179–92.

Halpern, A. R., and R.J. Zatorre. 1999. “When That Tune Runs Through Your Head: A PET Investigation of Auditory Imagery for Familiar Melodies.” *Cerebral Cortex* 9 (7): 697–704. doi:10.1093/cercor/9.7.697.

Halpern, A. R., R. J. Zatorre, M. Bouffard, and J. A. Johnson. 2004. “Behavioral and Neural Correlates of Perceived and Imagined Musical Timbre.” *Neuropsychologia* 42 (9): 1281–92. doi:10.1016/j.neuropsychologia.2003.12.017.

Hastie, Trevor. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer.

Haynes, John-Dylan, and Geraint Rees. 2005. “Predicting the Orientation of Invisible Stimuli from Activity in Human Primary Visual Cortex.” *Nature Neuroscience* 8 (5): 686–91. doi:10.1038/nn1445.

Herff, Christian, Dominic Heger, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. 2015. “Brain-to-Text: Decoding Spoken Phrases from Phone Representations in the Brain.” *Frontiers in Neuroscience*, June. doi:10.3389/fnins.2015.00217.

Herholz, Sibylle C., Claudia Lappe, Arne Knief, and Christo Pantev. 2008. “Neural Basis of Music Imagery and the Effect of Musical Expertise.” *The European Journal of Neuroscience* 28 (11): 2352–60. doi:10.1111/j.1460-9568.2008.06515.x.

Hickok, Gregory, Bradley Buchsbaum, Colin Humphries, and Tugan Muftuler. 2003. “Auditory-Motor Interaction Revealed by fMRI: Speech, Music, and Working Memory in Area Spt.” *Journal of Cognitive Neuroscience* 15 (5): 673–82. doi:10.1162/089892903322307393.

Hickok, Gregory, and David Poeppel. 2007. “The Cortical Organization of Speech Processing.” *Nature Reviews Neuroscience* 8 (5): 393–402. doi:10.1038/nrn2113.

Hinke, R M, X Hu, A E Stillman, S G Kim, H Merkle, R Salmi, and K Ugurbil. 1993. “Functional Magnetic Resonance Imaging of Broca’s Area during Internal Speech.” *Neuroreport* 4 (6): 675–78.

Hochberg, Leigh R., Daniel Bacher, Beata Jarosiewicz, Nicolas Y. Masse, John D. Simeral, Joern Vogel, Sami Haddadin, et al. 2012. “Reach and Grasp by People with Tetraplegia Using a Neurally Controlled Robotic Arm.” *Nature* 485 (7398): 372–75. doi:10.1038/nature11076.

Horikawa, T., M. Tamaki, Y. Miyawaki, and Y. Kamitani. 2013. “Neural Decoding of Visual Imagery During Sleep.” *Science* 340 (6132): 639–42. doi:10.1126/science.1234330.

Hotelling, Harold. 1940. “The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters.” *The Annals of Mathematical Statistics* 11 (3): 271–83. doi:10.1214/aoms/1177731867.

Houde, John F, and Edward F Chang. 2015. “The Cortical Computations Underlying Feedback Control in Vocal Production.” *Current Opinion in Neurobiology* 33 (August): 174–81. doi:10.1016/j.conb.2015.04.006.

Huang, Jie, Thomas H Carr, and Yue Cao. 2002. "Comparing Cortical Activations for Silent and Overt Speech Using Event-Related fMRI." *Human Brain Mapping* 15 (1): 39–53.

Hubbard, Timothy L. 2010. "Auditory Imagery: Empirical Findings." *Psychological Bulletin* 136 (2): 302–29. doi:10.1037/a0018436.

———. 2013. "Auditory Aspects of Auditory Imagery." In *Multisensory Imagery*, edited by Simon Lacey and Rebecca Lawson, 51–76. New York, NY: Springer New York. http://www.springerlink.com/index/10.1007/978-1-4614-5879-1_4.

Huth, Alexander G., Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. "Natural Speech Reveals the Semantic Maps That Tile Human Cerebral Cortex." *Nature* 532 (7600): 453–58. doi:10.1038/nature17637.

Ikeda, Shigeyuki, Tomohiro Shibata, Naoki Nakano, Rieko Okada, Naohiro Tsuyuguchi, Kazushi Ikeda, and Amami Kato. 2014. "Neural Decoding of Single Vowels during Covert Articulation Using Electrocorticography." *Frontiers in Human Neuroscience*, 125. doi:10.3389/fnhum.2014.00125.

Intons-Peterson, M J. 1980. "The Role of Loudness in Auditory Imagery." *Memory & Cognition* 8 (5): 385–93.

Intons-Peterson, MJ. 1992. "Components of Auditory Imagery." In *Auditory Imagery*, edited by Daniel Reisberg, 45–72. Hillsdale, N.J: L. Erlbaum Associates.

Janata, Petr, and Kaivon Paroo. 2006. "Acuity of Auditory Images in Pitch and Time." *Perception & Psychophysics* 68 (5): 829–44.

Kaneoke, Yoshiki, Tomohiro Donishi, Jun Iwatani, Satoshi Ukai, Kazuhiro Shinosaki, and Masaki Terada. 2012. "Variance and Autocorrelation of the Spontaneous Slow Brain Activity." Edited by Carles Soriano-Mas. *PLoS ONE* 7 (5): e38131. doi:10.1371/journal.pone.0038131.

Kaur, and Garg. 2012. "Speech Recognition System; Challenges and Techniques." *International Joural of Computer Science and Infomration Technlogies* 3.

Kennedy, J.F. 1961. "'Inaugural Address.'"

Khodagholy, Dion, Jennifer N Gelinas, Thomas Thesen, Werner Doyle, Orrin Devinsky, George G Malliaras, and György Buzsáki. 2014. "NeuroGrid: Recording Action Potentials from the Surface of the Brain." *Nature Neuroscience* 18 (2): 310–15. doi:10.1038/nn.3905.

Kosslyn, S. M., G. Ganis, and W. L. Thompson. 2001. "Neural Foundations of Imagery." *Nature Reviews. Neuroscience* 2 (9): 635–42. doi:10.1038/35090055.

Kosslyn, S. M., and W. L. Thompson. 2000. "Shared Mechanisms in Visual Imagery and Visual Perception: Insights from Cognitive Neuroscience." In *The New Cognitive Neurosciences*, edited by M. S. Gazzaniga, 2nd ed. Cambridge, MA: MIT Press.

———. n.d. "Shared Mechanisms in Visual Imagery and Visual Perception: Insights from Cognitive Neuroscience." *M. S. Gazzaniga (Ed.). The New Cognitive Neurosciences*, 2nd edition.

Kosslyn, Stephen M., and William L. Thompson. 2003. "When Is Early Visual Cortex Activated during Visual Mental Imagery?" *Psychological Bulletin* 129 (5): 723–46. doi:10.1037/0033-2909.129.5.723.

Kraemer, David J. M., C. Neil Macrae, Adam E. Green, and William M. Kelley. 2005. "Musical Imagery: Sound of Silence Activates Auditory Cortex." *Nature* 434 (7030): 158–158. doi:10.1038/434158a.

Kubanek, Jan, Peter Brunner, Aysegul Gunduz, David Poeppel, and Gerwin Schalk. 2013. "The Tracking of Speech Envelope in the Human Cortex." Edited by Antoni Rodriguez-Fornells. *PLoS ONE* 8 (1): e53398. doi:10.1371/journal.pone.0053398.

Lachaux, Jean-Philippe, Nikolai Axmacher, Florian Mormann, Eric Halgren, and Nathan E Crone. 2012a. "High-Frequency Neural Activity and Human Cognition: Past, Present and Possible Future of Intracranial EEG Research." *Progress in Neurobiology* 98 (3): 279–301. doi:10.1016/j.pneurobio.2012.06.008.

———. 2012b. "High-Frequency Neural Activity and Human Cognition: Past, Present and Possible Future of Intracranial EEG Research." *Progress in Neurobiology* 98 (3): 279–301. doi:10.1016/j.pneurobio.2012.06.008.

Lancaster, J L, M G Woldorff, L M Parsons, M Liotti, C S Freitas, L Rainey, P V Kochunov, D Nickerson, S A Mikiten, and P T Fox. 2000. "Automated Talairach Atlas Labels for Functional Brain Mapping." *Human Brain Mapping* 10 (3): 120–31.

Lartillot, Olivier, Petri Toiviainen, and Tuomas Eerola. 2008. "A Matlab Toolbox for Music Information Retrieval." In *Data Analysis, Machine Learning and Applications*, edited by Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, 261–68. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-540-78246-9_31.

Leonard, M. K., M.O. Baud, M.J. Sjerps, and E.F. Chang. 2016. "Perceptual Restoration of Masked Speech in Human Cortex." doi:10.1038/ncomms13619.

Leonard, M. K., K. E. Bouchard, C. Tang, and E. F. Chang. 2015. "Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex." *Journal of Neuroscience* 35 (18): 7203–14. doi:10.1523/JNEUROSCI.4100-14.2015.

Leuthardt, Eric C., Xiao-Mei Pei, Jonathan Breshears, Charles Gaona, Mohit Sharma, Zac Freudenberg, Dennis Barbour, and Gerwin Schalk. 2012. "Temporal Evolution of Gamma Activity in Human Cortex during an Overt and Covert Word Repetition Task." *Frontiers in Human Neuroscience* 6. doi:10.3389/fnhum.2012.00099.

Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran. 2004. "A Brain–computer Interface Using Electrocorticographic Signals in Humans." *Journal of Neural Engineering* 1 (2): 63–71. doi:10.1088/1741-2560/1/2/001.

Levelt. 1993. *Speaking: From Intention to Articulation*. Bradford Books,U.S. https://books.google.com/books?id=LbVCdCE-NQAC.

Llorens, A., Agnès Trébuchon, Catherine Liégeois-Chauvel, and F.-Xavier Alario. 2011. "Intra-Cranial Recordings of Brain Activity During Language Production." *Frontiers in Psychology*. doi:10.3389/fpsyg.2011.00375.

Lotte, Fabien, Jonathan S. Brumberg, Peter Brunner, Aysegul Gunduz, Anthony L. Ritaccio, Cuntai Guan, and Gerwin Schalk. 2015. "Electrocorticographic Representations of Segmental Features in Continuous Speech." *Frontiers in Human Neuroscience* 09 (February). doi:10.3389/fnhum.2015.00097.

Manning, J. R., J. Jacobs, I. Fried, and M. J. Kahana. 2009. "Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans." *Journal of Neuroscience* 29 (43): 13613–20. doi:10.1523/JNEUROSCI.2041-09.2009.

Martin, Stephanie, Peter Brunner, Chris Holdgraf, Hans-Jochen Heinze, Nathan E. Crone, Jochem Rieger, Gerwin Schalk, Robert T. Knight, and Brian N. Pasley. 2014. "Decoding Spectrotemporal Features of Overt and Covert Speech from the Human Cortex." *Frontiers in Neuroengineering*. doi:10.3389/fneng.2014.00014.

Martin, Stephanie, Peter Brunner, Iñaki Iturrate, José del R. Millán, Gerwin Schalk, Robert T. Knight, and Brian N. Pasley. 2016. "Word Pair Classification during Imagined Speech Using Direct Brain Recordings." *Scientific Reports* 6 (May): 25803. doi:10.1038/srep25803.

Martin, Stephanie, Christian Mikutta, Matthew K Leonard, Dylan Hungate, Stefan Koelsch, Edward F Chang, Jose del R. Millan, Robert T Knight, and Brian N Pasley. 2017. "Neural Encoding of Auditory Features during Music Perception and Imagery: Insight into the Brain of a Piano Player." *bioRxiv*, February. doi:10.1101/106617.

Mauchly, John W. 1940. "Significance Test for Sphericity of a Normal N-Variate Distribution." *The Annals of Mathematical Statistics* 11 (2): 204–9. doi:10.1214/aoms/1177731915.

McGuire, P K, D A Silbersweig, R M Murray, A S David, R S Frackowiak, and C D Frith. 1996. "Functional Anatomy of Inner Speech and Auditory Verbal Imagery." *Psychological Medicine* 26 (1): 29–38.

Meister, I. G., T. Krings, H. Foltys, B. Boroojerdi, M. Müller, R. Töpper, and A. Thron. 2004. "Playing Piano in the Mind--an fMRI Study on Music Imagery and Performance in Pianists." *Brain Research. Cognitive Brain Research* 19 (3): 219–28. doi:10.1016/j.cogbrainres.2003.12.005.

Mesgarani, and Edward F. Chang. 2012. "Selective Cortical Representation of Attended Speaker in Multi-Talker Speech Perception." *Nature* 485 (7397): 233–36. doi:10.1038/nature11020.

Mesgarani, N., C. Cheung, K. Johnson, and E. F. Chang. 2014. "Phonetic Feature Encoding in Human Superior Temporal Gyrus." *Science* 343 (6174): 1006–10. doi:10.1126/science.1245994.

Meyer, Martin, Stefan Elmer, Simon Baumann, and Lutz Jancke. 2007. "Short-Term Plasticity in the Auditory System: Differential Neural Responses to Perception and Imagery of Speech and Music." *Restorative Neurology and Neuroscience* 25 (3-4): 411–31.

Mikumo, Mariko. 1994. "Motor Encoding Strategy for Pitches of Melodies." *Music Perception: An Interdisciplinary Journal* 12 (2): 175–97. doi:10.2307/40285650.

Millán, F Galan, D Vanhooydonck, E Lew, J Philips, and M Nuttin. 2009. "Asynchronous Non-Invasive Brain-Actuated Control of an Intelligent Wheelchair." *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 2009: 3361–64. doi:10.1109/IEMBS.2009.5332828.

Miller, Kai J, Eric C Leuthardt, Gerwin Schalk, Rajesh P N Rao, Nicholas R Anderson, Daniel W Moran, John W Miller, and Jeffrey G Ojemann. 2007. "Spectral Changes in Cortical Surface Potentials during Motor Movement." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 27 (9): 2424–32. doi:10.1523/JNEUROSCI.3886-06.2007.

Minev, I. R., P. Musienko, A. Hirsch, Q. Barraud, N. Wenger, E. M. Moraud, J. Gandar, et al. 2015. "Electronic Dura Mater for Long-Term Multimodal Neural Interfaces." *Science* 347 (6218): 159–63. doi:10.1126/science.1260318.

Moses, David A, Nima Mesgarani, Matthew K Leonard, and Edward F Chang. 2016. "Neural Speech Recognition: Continuous Phoneme Decoding Using Spatiotemporal Representations of Human Cortical Activity." *Journal of Neural Engineering* 13 (5): 056004. doi:10.1088/1741-2560/13/5/056004.

"Mother Goose's Nursery Rhymes." 1867. 1877. A Collection of Alphabets, Rhymes, Tales and Jingles.

Mugler, Emily M, James L Patton, Robert D Flint, Zachary A Wright, Stephan U Schuele, Joshua Rosenow, Jerry J Shih, Dean J Krusienski, and Marc W Slutzky. 2014. "Direct Classification of All American English Phonemes Using Signals from Functional Speech Motor Cortex." *Journal of Neural Engineering* 11 (3): 035015. doi:10.1088/1741-2560/11/3/035015.

Nijboer, F., E. W. Sellers, J. Mellinger, M. A. Jordan, T. Matuz, A. Furdea, S. Halder, et al. 2008. "A P300-Based Brain-Computer Interface for People with Amyotrophic Lateral Sclerosis." *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 119 (8): 1909–16. doi:10.1016/j.clinph.2008.03.034.

Palmer, Erica D., Howard J. Rosen, Jeffrey G. Ojemann, Randy L. Buckner, William M. Kelley, and Steven E. Petersen. 2001. "An Event-Related fMRI Study of Overt and Covert Word Stem Completion." *NeuroImage* 14 (1): 182–93. doi:10.1006/nimg.2001.0779.

Pandarinath, Chethan, Paul Nuyujukian, Christine H Blabe, Brittany L Sorice, Jad Saab, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson. 2017. "High Performance Communication by People with Paralysis Using an Intracortical Brain-Computer Interface." *eLife* 6 (February). doi:10.7554/eLife.18554.

Pantev, C., R. Oostenveld, A. Engelien, B. Ross, L. E. Roberts, and M. Hoke. 1998. "Increased Auditory Cortical Representation in Musicians." *Nature* 392 (6678): 811–14. doi:10.1038/33918.

Partovi, Sasan, Florian Konrad, Sasan Karimi, Fabian Rengier, John K. Lyo, Lisa Zipp, Ernst Nennig, and Christoph Stippich. 2012. "Effects of Covert and Overt Paradigms in Clinical Language fMRI." *Academic Radiology* 19 (5): 518–25. doi:10.1016/j.acra.2011.12.017.

Pasley, Brian N., Stephen V. David, Nima Mesgarani, Adeen Flinker, Shihab A. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang. 2012. "Reconstructing Speech from Human Auditory Cortex." Edited by Robert Zatorre. *PLoS Biology* 10 (1): e1001251. doi:10.1371/journal.pbio.1001251.

Pasley, Brian N., and Robert T. Knight. 2013. "Decoding Speech for Understanding and Treating Aphasia." In *Progress in Brain Research*, 207:435–56. Elsevier. http://linkinghub.elsevier.com/retrieve/pii/B9780444633279000187.

Pei, Dennis L Barbour, Eric C Leuthardt, and Gerwin Schalk. 2011. "Decoding Vowels and Consonants in Spoken and Imagined Words Using Electrocorticographic Signals in Humans." *Journal of Neural Engineering*, no. 4 (August). doi:10.1088/1741-2560/8/4/046028.

Pei, Xiaomei, Eric C. Leuthardt, Charles M. Gaona, Peter Brunner, Jonathan R. Wolpaw, and Gerwin Schalk. 2011. "Spatiotemporal Dynamics of Electrocorticographic High Gamma Activity during Overt and Covert Word Repetition." *NeuroImage* 54 (4): 2960–72. doi:10.1016/j.neuroimage.2010.10.029.

Perdikis, S, R Leeb, J Williamson, A Ramsay, M Tavella, L Desideri, E-J Hoogerwerf, A Al-Khodairy, R Murray-Smith, and J d R Millán. 2014. "Clinical Evaluation of BrainTree, a Motor Imagery Hybrid BCI Speller." *Journal of Neural Engineering* 11 (3): 036003. doi:10.1088/1741-2560/11/3/036003.

Perrone-Bertolotti, M., L. Rapin, J.-P. Lachaux, M. Baciu, and H. Lœvenbruck. 2014. “What Is That Little Voice inside My Head? Inner Speech Phenomenology, Its Role in Cognitive Performance, and Its Relation to Self-Monitoring.” *Behavioural Brain Research* 261 (March): 220–39. doi:10.1016/j.bbr.2013.12.034.

Petsche, H., A. von Stein, and O. Filz. 1996. “EEG Aspects of Mentally Playing an Instrument.” *Brain Research. Cognitive Brain Research* 3 (2): 115–23.

Pitt, M A, and R G Crowder. 1992. “The Role of Spectral and Dynamic Cues in Imagery for Musical Timbre.” *Journal of Experimental Psychology. Human Perception and Performance* 18 (3): 728–38.

Price. 2000. “The Anatomy of Language: Contributions from Functional Neuroimaging.” *Journal of Anatomy* 197 (3): 335–59. doi:10.1046/j.1469-7580.2000.19730335.x.

Price, C. J. 2012. “A Review and Synthesis of the First 20years of PET and fMRI Studies of Heard Speech, Spoken Language and Reading.” *NeuroImage* 62 (2): 816–47. doi:10.1016/j.neuroimage.2012.04.062.

Pulvermuller, F., M. Huss, F. Kherif, F. Moscoso del Prado Martin, O. Hauk, and Y. Shtyrov. 2006. “Motor Cortex Maps Articulatory Features of Speech Sounds.” *Proceedings of the National Academy of Sciences* 103 (20): 7865–70. doi:10.1073/pnas.0509989103.

Rabiner, Lawrence R. 1993. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series. Englewood Cliffs, N.J: PTR Prentice Hall.

Rapp, B., S. Fischer-Baum, and M. Miozzo. 2015. “Modality and Morphology: What We Write May Not Be What We Say.” *Psychological Science* 26 (6): 892–902. doi:10.1177/0956797615573520.

Rauschecker, J. P. 2001. “Cortical Plasticity and Music.” *Annals of the New York Academy of Sciences* 930 (June): 330–36.

Reddy, Leila, Naotsugu Tsuchiya, and Thomas Serre. 2010. “Reading the Mind’s Eye: Decoding Category Information during Mental Imagery.” *NeuroImage* 50 (2): 818–25. doi:10.1016/j.neuroimage.2009.11.084.

Ritaccio, Anthony, Peter Brunner, Aysegul Gunduz, Dora Hermes, Lawrence J. Hirsch, Joshua Jacobs, Kyousuke Kamada, et al. 2014. “Proceedings of the Fifth International Workshop on Advances in Electrocorticography.” *Epilepsy & Behavior* 41 (December): 183–92. doi:10.1016/j.yebeh.2014.09.015.

Romanski, L. M., B. Tian, J. Fritz, M. Mishkin, P. S. Goldman-Rakic, and J. P. Rauschecker. 1999. “Dual Streams of Auditory Afferents Target Multiple Domains in the Primate Prefrontal Cortex.” *Nat Neurosci* 2 (12): 1131–36. doi:10.1038/16056.

Rosen, Howard J., Jeffrey G. Ojemann, John M. Ollinger, and Steve E. Petersen. 2000. “Comparison of Brain Activation during Word Retrieval Done Silently and Aloud Using fMRI.” *Brain and Cognition* 42 (2): 201–17. doi:10.1006/brcg.1999.1100.

Roth, M, J Decety, M Raybaudi, R Massarelli, C Delon-Martin, C Segebarth, S Morand, A Gemignani, M Décorps, and M Jeannerod. 1996. “Possible Involvement of Primary Motor Cortex in Mentally Simulated Movement: A Functional Magnetic Resonance Imaging Study.” *Neuroreport* 7 (7): 1280–84.

Rouse, A. G., J. J. Williams, J. J. Wheeler, and D. W. Moran. 2013. “Cortical Adaptation to a Chronic Micro-Electrocorticographic Brain Computer Interface.” *Journal of Neuroscience* 33 (4): 1326–30. doi:10.1523/JNEUROSCI.0271-12.2013.

Roy, E, and P Basler. 1955. “Lincoln, A. 1863. ‘The Gettysburg Address.’ In ‘The Collected Works of Abraham Lincoln.’” New Brunswick, NJ: Rutgers UP.

Saenz, Melissa, and Dave R.M. Langers. 2014. “Tonotopic Mapping of Human Auditory Cortex.” *Hearing Research* 307 (January): 42–52. doi:10.1016/j.heares.2013.07.016.

Sakoe, H., and S. Chiba. 1978. “Dynamic Programming Algorithm Optimization for Spoken Word Recognition.” *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26 (1): 43–49. doi:10.1109/TASSP.1978.1163055.

Schalk, G. 2010. *A Practical Guide to Brain-Computer Interfacing with BCI2000: General-Purpose Software for Brain-Computer Interface Research, Data Acquisition, Stimulus Presentation, and Brain Monitoring*. London ; New York: Springer.

Schalk, G, J Kubánek, K J Miller, N R Anderson, E C Leuthardt, J G Ojemann, D Limbrick, D Moran, L A Gerhardt, and J R Wolpaw. 2007. “Decoding Two-Dimensional Movement Trajectories Using Electrocorticographic Signals in Humans.” *Journal of Neural Engineering* 4 (3): 264–75. doi:10.1088/1741-2560/4/3/012.

Schalk, G., D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. 2004. “BCI2000: A General-Purpose Brain-Computer Interface (BCI) System.” *IEEE Transactions on Biomedical Engineering* 51 (6): 1034–43. doi:10.1109/TBME.2004.827072.

Schön, Daniele, Reyna Gordon, Aurélie Campagne, Cyrille Magne, Corine Astésano, Jean-Luc Anton, and Mireille Besson. 2010. “Similar Cerebral Networks in Language, Music and Song Perception.” *NeuroImage* 51 (1): 450–61. doi:10.1016/j.neuroimage.2010.02.023.

Schürmann, Martin, Tommi Raij, Nobuya Fujiki, and Riitta Hari. 2002. “Mind’s Ear in a Musician: Where and When in the Brain.” *NeuroImage* 16 (2): 434–40. doi:10.1006/nimg.2002.1098.

Shamma, Shihab. 2003. “Physiological Foundations of Temporal Integration in the Perception of Speech.” *Journal of Phonetics* 31 (3-4): 495–501. doi:10.1016/j.wocn.2003.09.001.

Shergill, S. S., E. T. Bullmore, M. J. Brammer, S. C. Williams, R. M. Murray, and P. K. McGuire. 2001. “A Functional Study of Auditory Verbal Imagery.” *Psychological Medicine* 31 (2): 241–53.

Shimodaira, Hiroshi, Ken-ichi Nom, Mitsuru Nakai, and Shigeki Sagayama. 2001. “Dynamic Time-Alignment Kernel in Support Vector Machine.” In , 921–28.

Shuster, Linda I, and Susan K Lemieux. 2005. “An fMRI Investigation of Covertly and Overtly Produced Mono- and Multisyllabic Words.” *Brain and Language* 93 (1): 20–31. doi:10.1016/j.bandl.2004.07.007.

Singleton, Nina Capone, and Brian B. Shulman, eds. 2014. *Language Development: Foundations, Processes, and Clinical Applications*. 2nd ed. Burlington, MA: Jones & Bartlett Learning.

Slutzky, Marc W, Luke R Jordan, Todd Krieg, Ming Chen, David J Mogul, and Lee E Miller. 2010. “Optimal Spacing of Surface Electrode Arrays for Brain–machine Interface Applications.” *Journal of Neural Engineering* 7 (2): 026004. doi:10.1088/1741-2560/7/2/026004.

Staba, Richard J., Charles L. Wilson, Anatol Bragin, Itzhak Fried, and Jerome Engel. 2002. “Quantitative Analysis of High-Frequency Oscillations (80-500 Hz) Recorded in Human Epileptic Hippocampus and Entorhinal Cortex.” *Journal of Neurophysiology* 88 (4): 1743–52.

Stanikov, A, C.F. Aliferis, D.P. Hardin, and I Guyon. 2011. “A Gentle Introduction to Support Vector Machines in Biomedicine, Volume 1: Theory and Methods.” In *A Gentle Introduction to Support Vector Machines in Biomedicine, Volume 1: Theory and Methods*. Singapore: Singapore: World Scientific Publishing Co. Pte. Ltd.

Stevenson, Richard J, and Trevor I Case. 2005. “Olfactory Imagery: A Review.” *Psychonomic Bulletin & Review* 12 (2): 244–64.

Sturm, Irene, Benjamin Blankertz, Cristhian Potes, Gerwin Schalk, and Gabriel Curio. 2014. “ECoG High Gamma Activity Reveals Distinct Cortical Representations of Lyrics Passages, Harmonic and Timbre-Related Changes in a Rock Song.” *Frontiers in Human Neuroscience* 8 (October). doi:10.3389/fnhum.2014.00798.

Tankus, Ariel, Itzhak Fried, and Shy Shoham. 2012. “Structured Neuronal Encoding and Decoding of Human Speech Features.” *Nature Communications* 3 (August): 1015. doi:10.1038/ncomms1995.

Theunissen, F.E., S.V. David, N.C. Singh, A. Hsu, W.E. Vinje, and J.L. Gallant. 2001. “Estimating Spatio-Temporal Receptive Fields of Auditory and Visual Neurons from Their Responses to Natural Stimuli.” *Network: Computation in Neural Systems* 12 (3): 289–316. doi:10.1080/net.12.3.289.316.

Theunissen, F. E., K. Sen, and A. J. Doupe. 2000. “Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 20 (6): 2315–31.

Thirion, Bertrand, Edouard Duschenay, Vincent Michel, Gael Varoquaux, Olivier Grisel, Jacob VanderPlas, alexandre granfort, fabian pedregosa, Andreas Mueller, and Gilles Louppe. 2011. “Scikitlearn.” *Journal of Machine Learning Research* 12: 2825–30.

Tian, B. 2004. “Processing of Frequency-Modulated Sounds in the Lateral Auditory Belt Cortex of the Rhesus Monkey.” *Journal of Neurophysiology* 92 (5): 2993–3013. doi:10.1152/jn.00472.2003.

Toga, Arthur W., and Paul M. Thompson. 2003. “Mapping Brain Asymmetry.” *Nature Reviews Neuroscience* 4 (1): 37–48. doi:10.1038/nrn1009.

Tomasello, Michael. 2008. *Origins of Human Communication*. The Jean Nicod Lectures 2008. Cambridge, Mass: MIT Press.

Towle, V. L., H.-A. Yoon, M. Castelle, J. C. Edgar, N. M. Biassou, D. M. Frim, J.-P. Spire, and M. H. Kohrman. 2008. “ECoG Gamma Activity during a Language Task: Differentiating Expressive and Receptive Speech Areas.” *Brain* 131 (8): 2013–27. doi:10.1093/brain/awn147.

Trask, R. L. 1999. *Language: The Basics*. 2nd ed. London ; New York: Routledge.

Vansteensel, Mariska J., Elmar G.M. Pels, Martin G. Bleichner, Mariana P. Branco, Timothy Denison, Zachary V. Freudenburg, Peter Gosselaar, et al. 2016. “Fully Implanted Brain–Computer Interface in a Locked-In Patient with ALS.” *New England Journal of Medicine* 375 (21): 2060–66. doi:10.1056/NEJMoa1608085.

Vaseghi, Saeed V. 2007. *Multimedia Signal Processing: Theory and Applications in Speech, Music and Communications*. Chichester, England ; Hoboken, NJ: J. Wiley.

Vitevitch, Michael S., Paul A. Luce, David B. Pisoni, and Edward T. Auer. 1999. “Phonotactics, Neighborhood Activation, and Lexical Access for Spoken Words.” *Brain and Language* 68 (1-2): 306–11. doi:10.1006/brln.1999.2116.

Waibel, Alex, and Kai-Fu Lee, eds. 1990. *Readings in Speech Recognition*. San Mateo, Calif: Morgan Kaufmann Publishers.

Wang, A. D. Degenhart, G. P. Sudre, D. A. Pomerleau, and E. C. Tyler-Kabara. 2011. “Decoding Semantic Information from Human Electrocorticographic (ECoG) Signals.” In , 6294–98. IEEE. doi:10.1109/IEMBS.2011.6091553.

Wang, Wei, Jennifer L. Collinger, Alan D. Degenhart, Elizabeth C. Tyler-Kabara, Andrew B. Schwartz, Daniel W. Moran, Douglas J. Weber, et al. 2013. “An Electrocorticographic Brain Interface in an Individual with Tetraplegia.” Edited by Shawn Hochman. *PLoS ONE* 8 (2): e55344. doi:10.1371/journal.pone.0055344.

Wilson, Stephen M, Ayşe Pinar Saygin, Martin I Sereno, and Marco Iacoboni. 2004. “Listening to Speech Activates Motor Areas Involved in Speech Production.” *Nature Neuroscience* 7 (7): 701–2. doi:10.1038/nn1263.

Winkler, István, Susan L. Denham, and Israel Nelken. 2009. “Modeling the Auditory Scene: Predictive Regularity Representations and Perceptual Objects.” *Trends in Cognitive Sciences* 13 (12): 532–40. doi:10.1016/j.tics.2009.09.003.

Wodlinger, B., A. D. Degenhart, J. L. Collinger, E. C. Tyler-Kabara, and Wei Wang. 2011. “The Impact of Electrode Characteristics on Electrocorticography (ECoG).” In , 3083–86. IEEE. doi:10.1109/IEMBS.2011.6090842.

Wolpaw, Jonathan R., Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. 2002. “Brain-Computer Interfaces for Communication and Control.” *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 113 (6): 767–91.

Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris. 1991. “An EEG-Based Brain-Computer Interface for Cursor Control.” *Electroencephalography and Clinical Neurophysiology* 78 (3): 252–59.

Wonnacott, Thomas H. 1990. *Introductory Statistics*. 5th ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Wu, Michael C.-K., Stephen V. David, and Jack L. Gallant. 2006. “Complete Functional Characterization of Sensory Neurons by System Identification.” *Annual Review of Neuroscience* 29: 477–505. doi:10.1146/annurev.neuro.29.051605.113024.

Yetkin, F. Z., T. A. Hammeke, S. J. Swanson, G. L. Morris, W. M. Mueller, T. L. McAuliffe, and V. M. Haughton. 1995. “A Comparison of Functional MR Activation Patterns during Silent and Audible Language Tasks.” *AJNR. American Journal of Neuroradiology* 16 (5): 1087–92.

Yoo, S S, C U Lee, and B G Choi. 2001. “Human Brain Mapping of Auditory Imagery: Event-Related Functional MRI Study.” *Neuroreport* 12 (14): 3045–49.

Zatorre, and Halpern. 1993. “Effect of Unilateral Temporal-Lobe Excision on Perception and Imagery of Songs.” *Neuropsychologia* 31 (3): 221–32.

Zatorre, Robert J., and Andrea R. Halpern. 2005. “Mental Concerts: Musical Imagery and Auditory Cortex.” *Neuron* 47 (1): 9–12. doi:10.1016/j.neuron.2005.06.013.

Zatorre, Robert J., Andrea R. Halpern, and Marc Bouffard. 2009. “Mental Reversal of Imagined Melodies: A Role for the Posterior Parietal Cortex.” *Journal of Cognitive Neuroscience* 22 (4): 775–89. doi:10.1162/jocn.2009.21239.

Zatorre, Robert J., Andrea R. Halpern, David W. Perry, Ernst Meyer, and Alan C. Evans. 1996. “Hearing in the Mind’s Ear: A PET Investigation of Musical Imagery and Perception.” *Journal of Cognitive Neuroscience* 8 (1): 29–46. doi:10.1162/jocn.1996.8.1.29.

Zion Golumbic, Elana M, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A Schevon, Guy M McKhann, Robert R Goodman, et al. 2013. “Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a ‘Cocktail Party.’” *Neuron* 77 (5): 980–91. doi:10.1016/j.neuron.2012.12.037.

STEPHANIE MARTIN

Rue du Simplon 12, 1006 Lausanne, Switzerland
✆ 0041 (0) 79 690 89 23 | ✉ steph.martin210@gmail.com | in stephanie-martin-673a1595 | DoB 24.10.1989

HIGHLIGHTS

•PhD in computational neuroscience – modelling language in the human brain.
•Neurobiology of language, machine learning, data mining, programming.
•Eager to expand leadership skills, maximize my potential and build connections.

PROFILE AND OBJECTIVES

PhD student currently finishing her thesis in computational neuroscience – with four years of experience in data analysis and machine learning for speech recognition from neural signals. Capable of designing, implementing and evaluating innovative research studies – by combining creativity, perseverance and problem solving skills. Excellent writing and communication skills. Eager to continuously develop new areas of expertise, and adapt well to any environment.

EDUCATION

2014-Now	PhD in Neuroengineering at Swiss Federal Institute of Technologies In collaboration with University of California, Berkeley
2012-2013	Master Thesis at University of California, Berkeley Knight Cognitive Neuroscience Lab (Grade: 6/6)
2011-2012	MSc in Bioengineering at Swiss Federal Institute of Technologies, Lausanne Specialization: Neuroscience; minor: Biomedical technologies (Grade: 5.6/6)
2008-2011	BSc in Life Sciences & Technology at Swiss Federal Institute of Technologies (Grade : 4.78/6)
2005-2008	Gymnase Sportif Auguste Piccard, Lausanne Biology and chemistry

CORE EXPERIENCE

2014-Now

PhD Thesis at *Brain-Machine Interface Lab – EPFL & UC Berkeley*
Subject: understand and decode imagined speech using intracranial recordings in the human brain.
Description: made critical contributions to the development of speech neuroprosthesis – by implementing various non-linear machine learning algorithms in order to model speech in the human cortex (e.g. support-vector machine, Hidden Markov Models, neural networks).

2012-2013

Master Thesis at *Knight Cognitive Neuroscience Lab – UC Berkeley (Grade: 6/6)*
Subject: reconstruct acoustic features of overt and imagined speech using electrocorticography.
Description: provided the first evidence of the ability to decode imagined speech – by reconstructing acoustic features from neural signals using regression models and dynamic time warping.

2012

Minor Project at *Chair in Non-invasive Brain and computer interface – EPFL* (Grade: 6/6)
Subject: eye artifact processing for single trial analysis of electroencephalogram.
Description: implemented algorithms, such as independent component analysis and regression, to remove eye artifacts from the electroencephalogram.

2008

Bachelor project at *Laboratory of molecular neurodegenerative research* – EPFL (Grade: 6/6)
Subject: The role of ATP13A2 in Parkinson's disease.
Description: investigated cell viability as a function of different gene expression, and a possible functional links between ATP13A2 and LRRK2.

TECHNICAL SKILLS

Programming	Matlab, Python, C/C++, R language, Latex
Engineering	Signal processing, data analysis, statistics, pattern classification, machine learning, data mining, electrical systems, brain-computer interaction.
Life science	Neuroscience, neurobiology of language processing, speech recognition, physiology, anatomy.

PROFESSIONAL EXPERIENCES

2014-Now	Teaching assistant for various classes, e.g. *Fundamentals of neural engineering, Data analysis and model classification* (Master courses – EPFL). Managed and organized exercise sessions, prepared and corrected semester projects and exams, supervised individual and groups of students.
2012	Internship at D-target, Medical Device CRO: organization and processes. Assisted in various clinical and regulatory affairs to bring medical devices to market.
2008-2012	Laboratory assistant at SICPA SA. Involved in quality assessment of security inks.

AWARDS AND PRICES

2016	Brain-Computer Interface Meeting Travel Award.
2015	Zeno-Karl Schindler Foundation Doctoral Exchange Grant – for collaborative inter-university doctoral top research in the fields of engineering.
2013	Mention of Excellence for the Master in Bioengineering
2013	Annaheim-Mattille Award (Fondation Marguerite) for an outstanding Master Project devoted to the rapprochement of Life Sciences and Informatics (bio-informatics and bio-inspired systems).
2013	Master Thesis Poster Award for the best poster of in Life Sciences and Technology.
2013	Social and Human Sciences (SHS) Award for an excellent first year SHS Master Project.

PUBLICATIONS

"Predictive models in cognitive electrophysiology – a practical guide." Holdgraf C., Martin S., Rieger J., Micheli C. 2017. (under revision at Frontiers in Systems Neuroscience)

Neural encoding of auditory features during music perception and imagery. Martin S., Mikutta C., Leonard M., Hungate D., Koelsch S., Chang E.F., Millán J. del R., Knight R.T., Pasley B.N. (under revision at PLOS Biology).

Word pair classification during imagined speech using direct brain recordings. Martin S., Brunner P., Iturrate I., Millán J. del R., Schalk G., Knight R.T., Pasley B.N., *Nature Scientific Report* (2016)

The use of intracranial recordings to decode human language: challenges and opportunities. Martin S., Millán J. del R., Knight R.T., Pasley B.N., *Brain and Language* (2016)

Understanding and decoding thoughts in the human brain. Martin S., Mikutta C., Knight R.T., Pasley B.N. *Frontiers for young minds* (2016)

Decoding spectrotemporal features of overt and covert speech from the human cortex. Martin, S., Brunner, P., Holdgraf, C., Heinze, H.J., Crone, N.E., Rieger J, Pasley, B.N., *Frontiers in Neuroengineering*. (2014)

LANGUAGES

French (native), English (fluent), German/Swiss German (fluent), Spanish (basics)

PERSONAL INTERESTS

Passion for travelling and discover new cultures, photography, climbing and mountaineering, ski touring and snowboard, sports, especially athletics (national and international competitions 1999-2010).