Aix*Marseille université
Socialement engagée

NNT : 0000AIXM0000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 2 Juillet 2024 par

Isaïh Mohamed

Lier la coordination verbale et les dynamiques neurales

Discipline
Biologie santé

Spécialité
Neuroscience

École doctorale
ED 62 – SCIENCES DE LA VIE ET DE LA SANTÉ

Laboratoire/Partenaires de recherche
Institut de Neurosciences des Systèmes
Équipe Dynamique de la Communication et des Processus Auditifs

Composition du jury

Dr. Stéphanie RIÈS Université de San Diego	**Rapportrice**
Pr. Sonja KOTZ Université de Maastricht	**Rapportrice et Présidente du jury**
Dr. Anaïs LLORENS Université Paris Cité et Université de Bourgogne Franche-Comté	**Membre invitée**
Dr. Daniele SCHÖN Aix-Marseille Université	**Directeur de thèse**
Dr. Leonardo LANCIA Aix-Marseille Université	**Co-directeur de thèse**

ILCB Institute of Language, Communication and the Brain

Institut de Neurosciences des Systèmes

NeuroSchool Aix*Marseille Université

Affidavit

Je soussigné, Isaïh Mohamed, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Daniele Schön et Leonardo Lancia, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille, le 5 mai 2024.



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Affidavit

I, undersigned, Isaïh Mohamed, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Daniele Schön et Leonardo Lancia, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Place Marseille, date May 5, 2024.



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Liste de publications et/ou brevets et participation aux conférences

1) Liste des publications et/ou brevet[1] réalisées dans le cadre du projet de thèse :

1. Hidalgo, C., **Mohamed, I.**, Zielinski, C., & Schön, D. (2022). The effect of speech degradation on the ability to track and predict turn structure in conversation. *Cortex*, *151*, 105-115.

2. **Mohamed, I. S.**, Mercier, M. R., Trebuchon, A., Morillon, B., Lancia, L., & Schon, D. (2024). Bridging verbal coordination and neural dynamics. *bioRxiv*, 2024-04. (soumis au journal Elife)

2) Participation aux conférences[2] et écoles d'été au cours de la période de thèse :

1. 3ème édition de l'École d'été organisée par l'Institute of Language Communication and the Brain (ILCB) (Août 2020)

2. Cuting EEG organisé par Maximilien Chaumon (Aix-en-Provence) – présentation de poster (Octobre 2021)

3. Journée des doctorants et post-doctorants (ILCB) Présentation conjointe « Le jeu de la conversation : repérer les dynamiques d'action conjointe dans le discours » (Juillet 2021)

4. Séminaire tutoré (ILCB) de Keith Doelling et Luc Arnal (Institut de l'Audition, Paris) – Présentation orale (Octobre 2021)

5. Workshop « Mind fair : a walk through language communication and brain" (ILCB) – démonstration de la tâche de mon projet de thèse (Mars 2022)

6. Conférence Alpine Brain Imaging Meeting (ABIM, Champéry – Suisse) organisé par l'université de Genève – présentation de poster (Janvier 2023)

7. Séminaire tutoré (ILCB) de Kate Watkins (Oxford university, Angleterre) – Présentation orale (Février 2023)

[1] Cette liste comprend les articles publiés, les articles soumis à publication et les articles en préparation ainsi que les livres, chapitres de livre et/ou toutes formes de valorisation des résultats des travaux propres à la discipline du projet de thèse. La référence aux publications doit suivre les règles standards de bibliographie et doit être conforme à la charte des publications d'AMU.

[2] Le terme « conférence » est générique. Il désigne à la fois « conférence », « congrès », « workshop », « colloques », « rencontres nationales et/ou internationales » … etc.

Indiquer si vous avez fait une présentation orale ou sous forme de poster.

8. Workshop “Time, Action and Language” organisé par Kate Watkins, Jennifer Coull, Marie-hélène Grosbras (Iméra, Marseille) – présentation orale (Mars 2023)

9. Workshop « Brain Rhythms and cortical computations” organisé par Keith Doelling, Luc Arnal, Anne-Lise Giraud, Benjamin Morillon (Institut de l’Audition, Paris) – présentation de poster (Avril 2023)

10. Colloques EDSVS – présentation d’un poster (2ème prix) (Mai 2023)

11. Conférence “Society for the Neurobiology of Language” – Présentation d’un poster (Octobre 2023)

12. Séminaire tutoré (ILCB) David Schlangen (Université de Potsdam, Allemagne) – présentation orale (Février 2024)

13. Journée des doctorants et post-doctorants « Kaleidoscope a collective journey through multidiscplinary insights » – démonstration de la tâche de mon projet de thèse (Mars 2024)

Résumé

Notre utilisation du langage est profondément sociale par nature, se déroule essentiellement dans des contextes d'interaction et est façonnée par des dynamiques de coordination précises que les interlocuteurs doivent respecter. Ainsi, le langage en interaction exige un ajustement rapide de la production de la parole.

Dans cette thèse, l'objectif est de lier la coordination verbale et les dynamiques neurales. Pour cela, nous avons développé un partenaire virtuel fondé sur un modèle à oscillateurs couplés en temps réel qui permet de moduler sa capacité à synchroniser sa parole avec celle d'un locuteur. Ceci conduit à une adaptation mutuelle dynamique des productions verbales. Nous avons ensuite enregistré l'activité cérébrale intracrânienne de 16 patients atteints d'épilepsie pharmaco-résistante pendant qu'ils effectuaient une tâche de coordination verbale avec le partenaire virtuel (VP). En l'occurence, les patients devaient répéter de courtes phrases de manière synchrone avec le partenaire virtuel. Cette tâche de parole synchronisée interactive est efficace pour mettre en évidence les voies dorsales et ventrales du langage. Deuxièmement, la combinaison de la coordination verbale et de l'activité neuronale résolue dans le temps montre des motifs d'activité plus spatialement différenciés et différents types de sensibilité neuronale le long de la voie dorsale, des activités basse fréquence jusqu'aux activités à haute fréquence. Plus précisément, quatre régions d'intérêt dans l'hémisphère gauche montrent une forte implication, à savoir le gyrus temporal supérieur (STG BA41/42 et BA22), le lobule pariétal inférieur (IPL BA 40) et le gyrus frontal inférieur (IFG BA44). Troisièmement, l'activité à haute fréquence dans les régions auditives secondaires est très sensible à la dynamique de coordination verbale, alors que les régions primaires ne le sont pas. Enfin, l'activité à haute fréquence de l'IFG BA44 indexe spécifiquement les ajustements de coordination qui sont continuellement nécessaires pour compenser la déviation de la synchronisation. Ce dernier point souligne la relation entre les capacités de coordination et les circuits anatomiques de l'IFG ainsi que sa dynamique neurale montrant qu'elle joue un rôle important dans les ajustements temporels de la parole qui sont nécessaires pour se synchroniser à un signal de parole externe.

Nos résultats obtenus grâce à la tâche de parole synchronisée interactive soulignent une nécessité pour les modèles du contrôle moteur de la parole de tenir compte des influences des entrées de parole externe sur les processus sensorimoteurs pendant la production de la parole. De plus, l'utilisation de la parole synchronisée peut offrir une contribution dans la recherche clinique concernant la cartographie fonctionnelle de la production du langage, en particulier sur les processus de coordination de la parole et ainsi réduire les risques postopératoires de développer des troubles du langage.

Mots-clés : coordination verbale, adaptation mutuelle, dynamique neurale, perception/production de la parole, gyrus frontal inférieur.

Abstract

Language is one of the richest and most complex functions in human beings. Our use of language, which is profoundly social in nature, essentially takes place in interactive contexts and is shaped by precise coordination dynamics that interlocutors must observe. Thus, language interaction is high demanding on fast adjustment of speech production.

In the present PhD thesis, we aim to bridge verbal coordination and neural dynamics. In this purpose, we developed a real-time coupled-oscillators virtual partner that allows – by changing the coupling strength parameters – to modulate its ability to synchronize speech with a speaker. This leading to a dynamic mutual adaptation of verbal productions. Then, we recorded the intracranial brain activity of 16 patients with drug-resistant epilepsy while they performed a verbal coordination task with the virtual partner (VP). More precisely, patients had to repeat short sentences synchronously with the VP. This interactive synchronous speech task is efficient to highlight both the dorsal and ventral language pathways. Secondly, combining time-resolved verbal coordination and neural activity shows more spatially differentiated patterns and different types of neural sensitivity along the dorsal pathway from low to high frequency activity. More precisely, four regions of interest in the left hemisphere show strong involvement, namely the superior temporal gyrus (STG BA41/42 and BA22), the inferior parietal lobule (IPL BA 40) and the inferior frontal gyrus (IFG BA44). Thirdly, high-frequency activity in secondary auditory regions is highly sensitive to verbal coordinative dynamics, while primary regions are not. Finally, the high frequency activity of the IFG BA44 specifically indexes the online coordinative adjustments that are continuously required to compensate the deviation from synchronization. This latter point emphasizes the relation between coordination abilities and anatomical circuitry of the IFG to the neural dynamics of this same region, showing that it plays an important role in the temporal adjustment of speech that are necessary to synchronize to external speech.

Overall, our results obtained through the interactive synchronous speech task highlight a necessity for speech motor control models to incorporates the influences of external speech inputs on sensorimotor processes during speech. Moreover, the use of synchronous speech could offer a new avenue in clinical research to contribute to the functional mapping concerning speech production, more specifically speech coordination processes and thus reduce postoperative risks of developing language impairments.

Keywords : verbal coordination, mutual adaptation, neural dynamics, speech perception/production, inferior frontal gyrus

Remerciements

Mes premiers remerciements sont adressés à mes parents. Maman, Papa, pour avoir toujours cru en moi, pour votre soutien et votre amour, toujours.

Je tiens à remercier toutes les enseignantes et enseignants que j'ai eu la chance de rencontrer durant l'ensemble de mon parcours scolaire. De l'école primaire jusqu'aux bancs de l'université, pour m'avoir appris à lire, écrire, compter, raisonner, pour m'avoir appris à apprendre.

Je remercie mon directeur de thèse Daniele Schön, pour la confiance qu'il m'a accordée, pour m'avoir toujours poussé à persévérer peu importe les difficultés, pour ses conseils toujours indispensables, pour son temps et son écoute. Je n'aurais pas pu rêver meilleure personne pour m'accompagner durant ces années.

Je remercie mon co-directeur de thèse Leonardo Lancia, pour son énergie et sa joie de vivre, pour tout le temps qu'il a consacré pour m'aider durant ce projet. Pour ses précieux conseils et m'avoir remotivé quand il le fallait. Rien n'aurait pu être possible sans lui.

Pour m'avoir accueilli dès mon stage de master, m'avoir fait apprécier la vie d'équipe et beaucoup fait rire, pour avoir toujours pris un peu de votre temps quand j'en avais besoin, Jacques, Arnaud, Jérémy un grand merci.

Je remercie Emma et Aurélie, pour être de formidables collègues de bureau, pour tous les agréables moments passés ensemble : courage !! (N'oublie pas de ranger ton bureau Emma)

Pour toutes les intéressantes conversations, les balades de fin de journée, ton écoute et ton amitié, je remercie chaleureusement Paul.

À l'ensemble des (post) doctorantes et doctorants que j'ai pu rencontrer dans le laboratoire et les agréables moments passés, je remercie Céline, Noémie, Akanksha, Bissera, Nadège, Anne, Laure, Chiara, Camille, Julie, Maude, Hiro, Victor, Pierre et enfin Clément (mention spéciale pour toutes ses passions débordantes !).

Je remercie Théo, Colas, Thibaut et Flavien pour les très bons moments passés depuis notre master jusqu'à aujourd'hui et pour avoir toujours su me remotiver quand j'en avais besoin !

Merci à Rayane, toujours présent malgré le temps.

Enfin, pour les petites variations. Toujours.

les petites variations.





Ce sont nos douces syntonies et ces rires qui nous prennent.
Nos regards qui se croisent, et s'embrassent sans prévenir.
Le mauve sur tes ongles, le rouge sur tes joues,
L'encre autour de tes yeux, un océan entier de toi.



La marche de nos ombres qui décore les murs la nuit,
Et nos tendresses, silencieuses, qui s'échangent à nos au revoir.
Quelques photos et souvenirs, pour enfermer les cendres de nos vies.
Toute une éternité en quelques instants, et les traces des autres en nous.



Ce sont des galeries sans fin, des lumières qui aveuglent,
Et parfois des murmures d'oiseaux dans le ciel.
Une infinité de tableaux défilant sous nos yeux,
Et souvent des bouquets de fleurs pour réchauffer nos cœurs.




Et si le cœur bat, c'est peut-être pour celui des autres aussi.

Ce sont nos voix qui se perdent dans les salles,
Les sons des guitares et des batteries qui pansent nos peines.
Ces perles sur ton cou, tes monstres irrésistibles qui dansent derrière ma poitrine.
Les échos de ton pouls qui résonnent dans le creux de ma main.
Tout l'univers, l'horreur, l'infini sur ta peau, qui déteint sur la mienne.




Ce sont nos bras qui s'enlacent, nos esprits tordus et nos mots qui s'amusent.
Nos grimaces, nos âmes vulnérables, et tous nos crimes d'affection.

C'est le vent qui se lève, le feu et la foudre dans les cieux quand nos têtes se perdent.
Ce sont ces frissons quand la peur revient, ces frissons, toujours plus forts quand l'affection nous tient.
Cette frénésie qui gronde et nos cœurs qui s'emballent quand « Disco » sonnent ses dernières notes.



C'est la vie, nos vies, les jours qui se jouent de nous.

Et encore nos rires, toujours, car il n'y a qu'eux qui comptent.
Ce ne sont que des pensées et des sentiments qui ne disent jamais leurs noms.



Et parfois des larmes sur nos joues.
Car parfois, il est douloureux d'être heureux.

Table of contents

Affidavit **2**

Résumé **6**

Abstract **7**

Remerciements 8

Chapter 1 – General introduction **12**

I. The study of language in Neuroscience 13
1. The Perception 13
a. Low-level processing and primary auditory cortex 13
b. Speech processing in auditory cortex 14
i. Caracteristics of the speech signal 14
ii. The role of superior temporal gyrus in speech processing 16
c. Towards more naturalistic stimuli in speech perception studies 21
2. The production 22
a. Investigations and modelling of speech production 22
i. How speech production can be studied? 22
ii. Lemma model of lexcial selection and form encoding 23
iii. Feedback control and predictive coding framework 25
b. The case of auditory feedback 27
c. Focus on intracranial recordings 29
i. The role of sensorimotor cortex in speech production 29
ii. Speech planning 31

II. The language in interaction 33
1. Conversation as joint activity 33
a. Joint activity 33
b. The case of conversation 35
c. Perception and production in conversation 39
2. Convergence and interactive alignment 39
3. Introducing interactivity into experimental paradigms 42
a. Turn-taking paradigm 42
b. The case of hyperscanning 48
i. Principle and method 48
ii. The use of hyperscanning to study language in interaction 49
c. The case of synchronous speech 53

Rationale / transition 56

Chapter 2 – An interactive synchronous speech task **57**
1. How to study verbal coordination in interactive context? 58
2. The behavioural study 58
a. Materials and methods 59
b. The virtual partner 62
c. Results 63
3. Adaptation of the task in a clinical environment 67

Chapter 3 – Bridging verbal coordination and neural dynamics. **70**

Chapter 4 – Clinical perspectives **97**

1. Clinical investigations of drug-resistant epilepsy....98
2. Clinical functional mapping of cognitive functions....99
   a. Pure tones and BaPa....100
   b. Picture naming....102
   c. Electrical stimulation and Wada test....103
3. Bridging fundamental research to clinical results....104
   a. Comparisons with passive listening (Ba/Pa) and electrical stimulation....104
   b. Comparisons with picture naming....105

Chapter 5 – General discussion....112
1. Summary of the results....113
2. The contribution of synchronous speech to speech motor control models....114
3. Limitations and future directions....116
4. About language in neuroscience research....118

Conclusion....126

Bibliography....**127**

Appendix – The effect of speech degradation on the ability to track and predict turn structure in conversation....145

Chapter 1 – General introduction

I. The study of language in Neuroscience

The study of language has been a central research topic for several decades, spanning various disciplines such as psycholinguistics, neuropsychology, and particularly neurophysiology, each allowing to enrich our knowledge of how language is processed and represented by the brain. Initially, the study of language in neuroscience was done by separating its two essential aspects, namely its perception and its production, then by gradually reconnecting them. Advances in neuroimaging techniques as well as in techniques for recording neural activity have made it possible to measure brain responses with greater fidelity and better spatio-temporal resolution, the means and methods of analyzing the recorded signals following the same progression trajectory made it possible to analyze larger, richer and more complex data sets. Here attention will be paid to investigations of the perception and production of language in an isolated context and then to the close relationships linking them. In addition, particular attention will be paid to the nature and degree of ecology of the stimuli and tasks used, as well as investigations based on intracranial recordings to study and measure the neural correlates of language.

1. The Perception

a. Low-level processing and primary auditory cortex

The basic processing of any type of sound in the human brain, including speech, begins with the vibration of the tympanic membrane caused by sound pressure waves, this resulting in the movement of the ossicular chain of the middle ear which is a three-bone system that acts as a lever to transmit sound information to the basilar membrane of the cochlea in the inner ear. Reacting to sound frequencies in a topographic manner, there is frequency selectivity occurring along the membrane, its base responding preferentially to high frequencies and its apex to low frequencies. This so-called tonotopic organization is preserved when auditory information is processed and transmitted from the inner ear to the brainstem, passing from the cochlear nucleus to the superior olive, to the medial lemniscus and the inferior colliculus, to the medial geniculate nucleus (MGN) of the thalamus and finally to the primary auditory cortex, located in humans at the level of Heschl's gyrus (Pickles, 2015). This process represents the very first so-called low-level processing of auditory perception.

The most commonly used classical paradigms in hearing studies focus on using passive listening tasks. These consist of listening to different types of sound while remaining still. Typically, a passive listening task involves presenting auditory stimuli simultaneously to both ears using earphones or headphones and asking participants to passively listen to the acoustic stimuli. This most often includes explicit instructions inviting them to remain still, sometimes with their eyes closed or to fix a target on a screen

in front of them. So, among the different stimuli that can be presented we can count the use of pure tones or white noise. Artificially created, several copies are generated varying in their frequency, their duration or even their intensity (Jäncke et al., 2002; Liégeois-Chauvel et al., 2004, Trébuchon et al., 2021). Thus, numerous studies have been able to demonstrate that the primary auditory cortex, particularly Heschl's gyrus, also presents a tonotopic organization (Howard et al., 1996; Formisano et al., 2003; Saenz et al., 2014; Nourski, 2017; Hamilton et al., 2021). Pure tones, although very useful to define and serve as a functional localizer at the level of the primary auditory cortex (Trébuchon et al., 2005 ; Trébuchon et al., 2021), are not natural stimuli and are far from representing language. Despite everything, we can also cite the use of other stimuli that are a little closer to ecological sounds such as sounds of natural environments (rain, wind, lightning), animal vocalizations (monkey, lion, horse), noises of tools (keys, scissors, vacuum cleaner) or even human vocalizations without it being language strictly speaking (baby crying, laughter, coughing sounds) (Moerel et al. 2012).

b. Speech processing in auditory cortex

i. Caracteristics of the speech signal

Human speech represents a particular acoustic signal, taking its origin from the flow of air coming from the lungs, crossing the vocal cords which open and close rapidly in a quasi-periodic manner (this frequency represents the fundamental frequency, noted F0), it results sound energy. The complex manipulation of this airflow through active variations in the shape of the vocal tract using the coordination of otolaryngeal structures such as the palate, tongue and lips allows the construction of a wide variety of distinctive sound units. We thus distinguish consonants and vowels as categories of speech sounds that differ acoustically and are described by distinct sets of articulatory characteristics. For consonants the airflow through the vocal tract is partially obstructed and they are characterised by the location (place of articulation) and the way in which the airflow through the mouth is restricted by the articulators (mode articulation). The two main categories being plosives and fricatives. Plosives are produced by first blocking the flow of air at the mouth, pharynx or glottis then suddenly releasing the air (e.g. /p/, /b/, /t/, /k/). Fricatives are produced by partially blocking the flow of air in the mouth (e.g. /f/, /v/, /s/, /z/). Vowels are characterised by an open, unobstructed configuration of the upper vocal cavity, which is shaped by the position of the tongue in the mouth and the arrangement of the lips during sound production. Natural language is then structured around consonants and vowels which appear in the form of highly organised repeated patterns which define the syllables. When syllables are strung together in longer utterances, this alternating structure of consonants and vowels contributes to what is considered the rhythm of speech. The peaks and valleys of these speech intensity modulations constitute the speech amplitude envelope that is captured in low frequencies and is essential for speech intelligibility (Drullman et al., 1994 ; Oxenham et al., 2017).

Consequently, speech sounds can be described by the acoustic properties of the sound wave or by the articulatory properties linked to the way in which the speech sound was formed from the vocal tract to its articulation. Moreover, these properties are subject to vary within the same individual and between individuals, a phenomenon defined as categorization (see Figure 1). Ultimately, the assembly of consonants and vowels allows the formation of syllables, the latter allowing the creation of words and the assembly of these same words allowing the construction of sentences (Hickok, 2012). Speech perception is thus based on a set of transformations which convert a complex acoustic signal into discrete and interpretable linguistic units. So-called language stimuli can then be considered from the moment there is use of a meaningful speech signal, namely syllables, words or sentences.

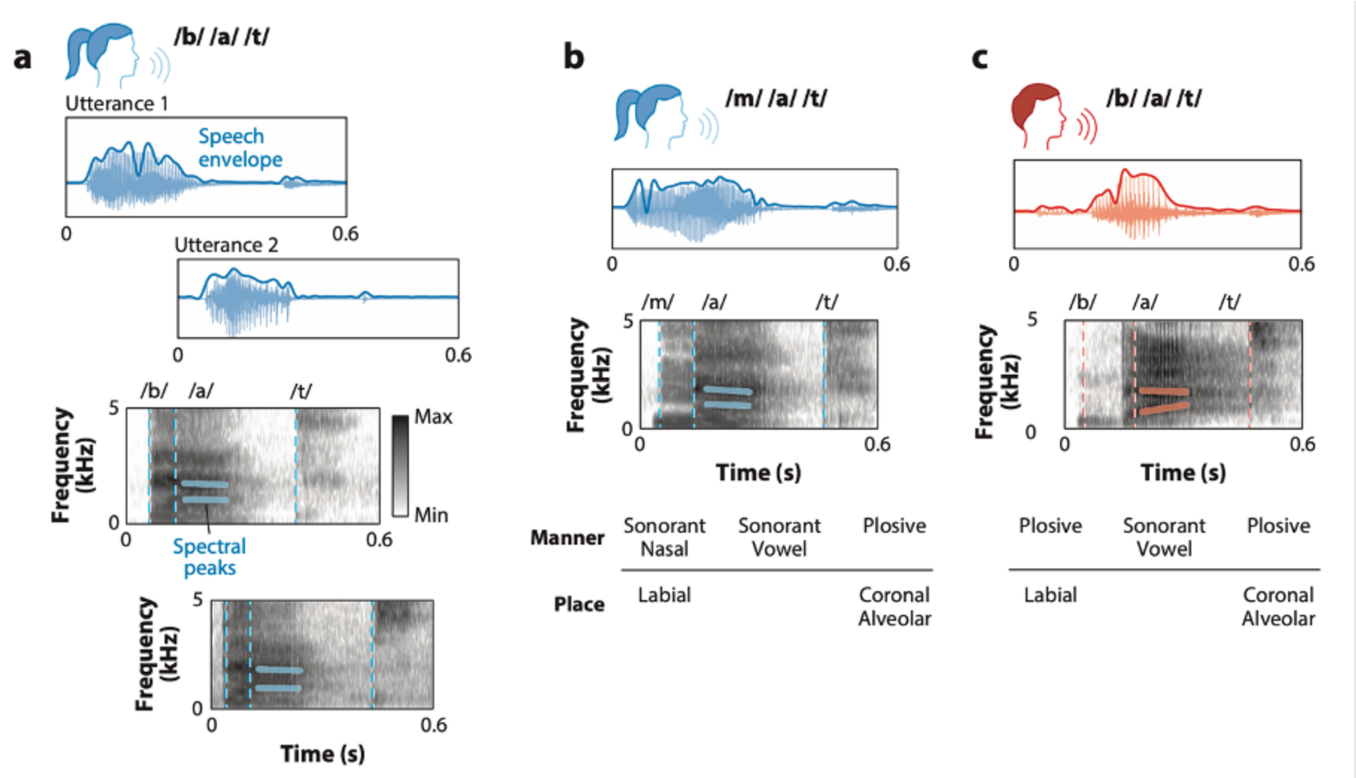


Figure 1 – Within and between speaker variability in acoustic and articulatory properties of speech signal. (a) A higher-pitch speaker produces two instances of "bat" slightly differently (labeled as utterance 1 and utterance 2), but both speech sequences map onto the same linguistic content. Key within-speaker differences in the speech waveform and spectrogram representation of the acoustic signal include changes in the amplitude of the speech envelope, shifted spectral peaks, and different final phoneme durations. **(b)** The same speaker as in panel a produces the word "mat." Corresponding acoustic-phonetic features are shown in the lowest panel, indicating the manner and place of the articulatory gesture that produces the corresponding sound. **(c)** A different, lower-pitch speaker than the speakers in panels a and b produces the word "bat." Key between-speaker differences in the speech waveform and spectrogram representation of the acoustic signal include changes in the amplitude of the speech envelope and shifted spectral peaks. Between-speaker variability can be due to several

specific speaker characteristics, such as the length of the speaker's vocal tract, speaker rate, and accent. *From Bhaya-Grossman & Chang, 2022*.

ii. The role of superior temporal gyrus in speech processing

The early anatomical studies on lesions concerning the temporal lobe (Wernicke, 1874; Geschwind, 1970) already suggested the superior temporal gyrus (STG) as a critical location for speech perception and language comprehension. Specifically, the STG contains part of the secondary auditory cortex by encompassing Brodmann area 22, the lateral parts of Brodmann areas 41 and 42 as well as Broadmann area 38 (Friederici, 2011 ; see Figure 2A). The STG is also defined more simply with a medial part and an anterior part (Hamilton et al., 2021 ; see Figure 2B). Numerous studies on patients manifesting lesions (ischemic strokes) in these regions (generally on the left hemisphere), particularly at the posterior level of the superior temporal gyrus (pSTG), have been able to show that acute dysfunctions were associated with an alteration in the auditory comprehension of words, or that reperfusion of the pSTG restored this comprehension and that the severity of the lesions was correlated with the severity of the alterations in the words comprehension (Hillis et al., 2017).

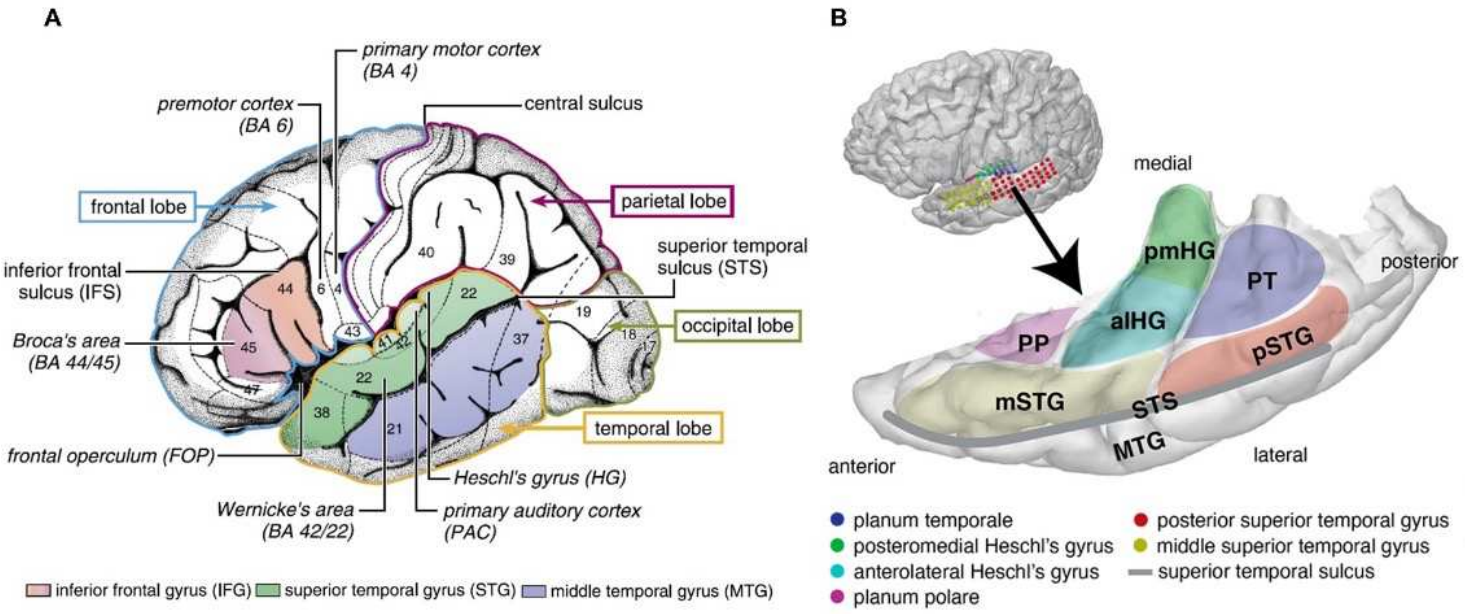


Figure 2 – Anatomy of the auditory cortex and parcellations of the superior temporal gyrus. (A) Anatomical and cytoarchitectonic details of the left hemisphere. Numbers indicate language-relevant Brodmann Areas (BA) which Brodmann (1909) defined on the basis of cytoarchitectonic characteristics. **(B)** Top left shows a three-dimensional reconstruction of the left hemisphere cerebrum of a patient with electrode recordings (grids) over the auditory cortex. Bottom right shows an enlarged temporal lobe, with frontal, parietal, and occipital lobes removed for ease of viewing. The position of the primary auditory cortex is thought to be located near the posteromedial Heschl's

gyrus (pmHG). Other regions of the temporal plane, located deep within the Sylvian fissure, include the planum temporale (PT), posteromedial andanterolateral Heschl's gyrus (pm/al HG), and planum polare (PP) most anteriorly. The lateral surface of the temporal lobe, just ventral from the temporal plane, includes the posterior and mid-superior temporal gyrus (STG). Ventrally to the STG is the middle temporal gyrus (MTG), which is separated from the STG by the superior temporal sulcus (STS). *Adapted from Friederici, 2011 (panel A, it's important to precise the absence of the basal ganglia and cerebellum) and Hamilton et al., 2021 (panel B).*

The use of different types of speech stimuli (syllables, isolated words and sentences) as well as technological advances allowing the dynamic measurement of brain activity at high resolution, such as recordings at the intracranial level, have made it possible to determine which elements of speech are represented by neuronal activity and according to what criteria. Intracranial data are generally collected from patients contracting drug-resistant epilepsies and where the functional exploration of major cognitive functions (e.g. language, memory) is essential before carrying out any attempt of surgery to eliminate the area containing the epileptogenic zone. This then involves the application of electrodes at the intracranial level according to different methods such as electrocorticography (ECoG) which consists of the installation of a grid of electrodes covering a surface of the cortex or the implantation of electrodes in the deep brain structures (stereoencephalography, sEEG) (Mercier et al., 2022). This type of data makes it possible to investigate high-frequency neuronal activity or more commonly called high-gamma activity (HGA: 70-150 Hz) which reflects very local neuronal spiking activity and the dendritic activity of several thousand neurons. (Steinschneider et al. 2008 ; Leszczyński et al., 2020). Thus, studies using ECoG recordings and others using recordings at the level of single neurons have demonstrated that neuronal populations included in the STG are particularly sensitive to the mode of articulation and respond selectively to acoustic and phonetic characteristics (e.g. plosive, fricative, vowel) of the speech signal (Mesgarani et al. 2014 ; Hamilton et al., 2018 ; Chan et al., 2014 ; Lakertz et al., 2021 ; Leonard et al., 2024). The structure of the speech signal, unfolding over time and consisting of sequences of elements (e.g. syllables -> words -> sentences), is a fundamental aspect in communication. This temporal aspect of speech was also investigated, and the STG also showed its involvement on very specific points. By having patients passively listen to simple sentences (e.g. "Have you got enough blankets?"; "It had gone like clockwork."), two very distinct populations of responses (high-gamma activity) at the level of the STG were identified, a first population preferentially located at the level of the pSTG selective to the onset of speech (particularly following at least 200 ms of silence) and a second rather localised at the anterior level characterised by sustained response activity throughout the duration of the sentence (Hullet et al., 2016; Hamilton et al., 2018). Sentence onsets play an essential role in the segmentation of the speech signal, making it possible to delimit sentence units. However stimuli of this type such as isolated sentences are unnatural, lacking both context and spontaneity. Hamilton and al. (2018) used natural speech to investigate if these responses were similar in a more ecological context. Specifically, in their case it was

extracts of natural speech from a movie (Forrest Gump by Robert Zemeckis) which were read by different speakers (woman and man). The same results were found when listening to these extracts of narration and dialogue (see Figure 3),the gradient of the two populations extending from the posterior parts to the more anterior parts of the STG, the responses to onset localizing at the level of the pSTG.

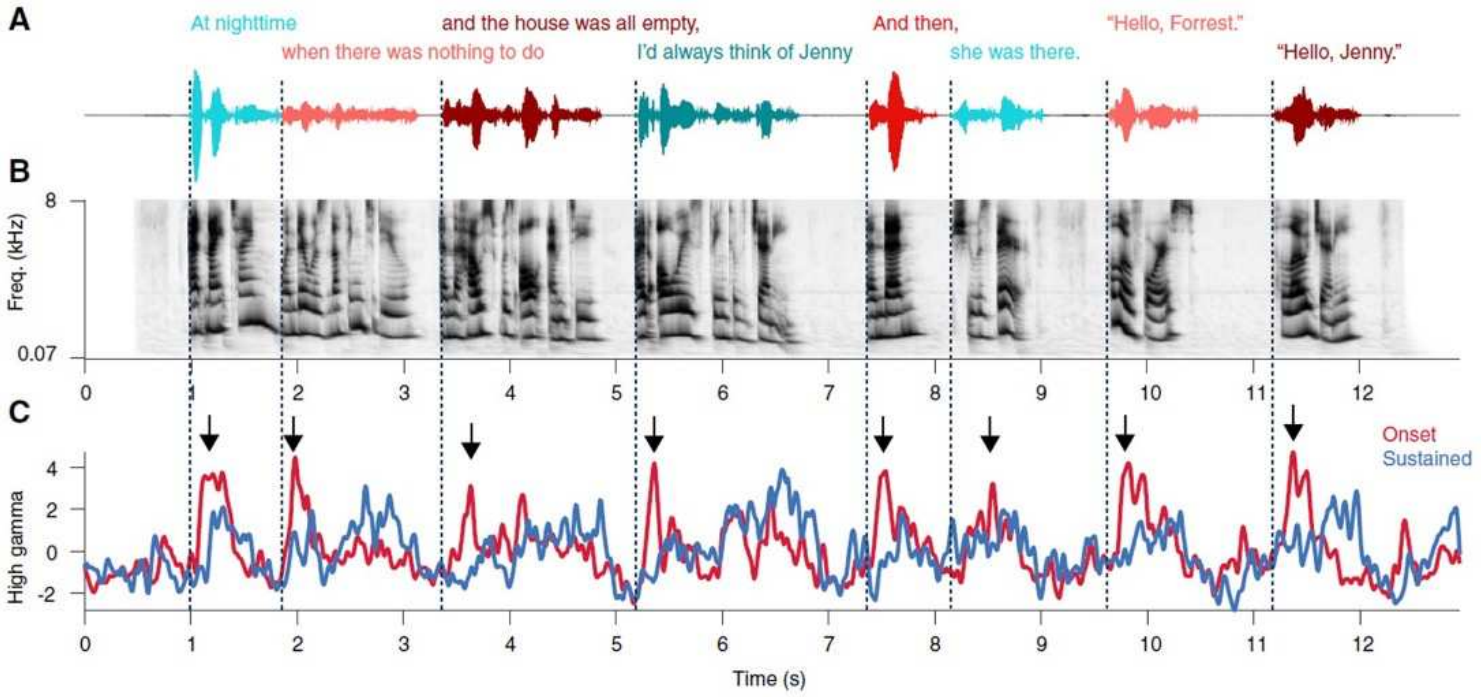


Figure 3 – Onset and sustained responses in natural continuous speech. (A) Sound waveform and transcription from a movie. The stimulus contains multiple sentences and phrases, with natural pauses in between utterances. Parts of the sentence are shown in color for ease of alignment with the sentence waveforms. **(B)** Spectrogram of the example stimulus shown in (A). **(C)** Example activity for an onset (red) and a sustained electrode (blue) from a single patient. The onset electrode activity demarcates acoustic cues for sentence and phrase-level pauses, such as after ‘‘at nighttime’’ or after ‘‘all empty’’. Longer pauses, such as those in between the dialog, also elicit strong onset responses from this electrode. *Adapted from Hamilton et al. 2018*.

Another crucial component of the speech signal is also processed within the auditory cortex, particularly at the level of the STG. Numerous studies in the literature demonstrate that brain activity continually tracks the speech envelope. The study conducted by Luo & Poeppel (2007) revealed that the pattern of phase responses in the theta frequency band (4-8 Hz), recorded in the auditory cortex using magnetoencephalography (MEG), enables the tracking and discrmination of spoken sentences. Furthermore, this ability to discriminate sentences correlates with speech intelligibility. This subsequently contributed to the development of the neural oscillations framework (Obleser & Kayser, 2019 for a review) highlighting that during speech

perception there is a coupling between theta and gamma bands in the auditory cortex. The phase of slow rhythmic activity would directly reflect the gain (e.g., excitability, spiking activity) of local neurons (see Figure 4; Giraud & Poeppel, 2012; Morillon et al., 2012; Hyafil et al., 2015; Lizarazu et al. 2019).

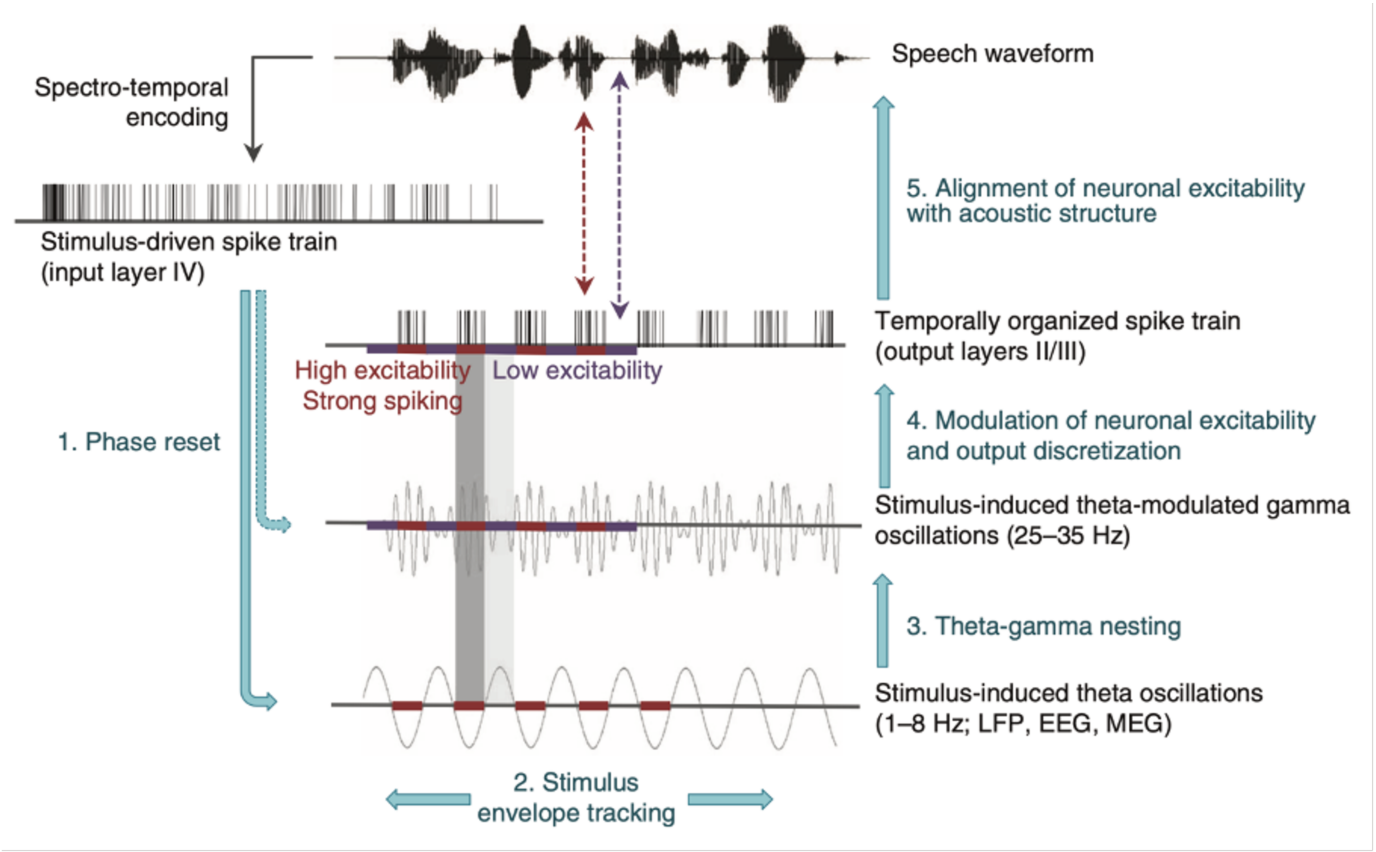


Figure 4 – A theory of early oscillation-based operations in speech perception. Five operations allow connected speech to be parsed by cortical theta and gamma oscillations. We assume a high- resolution spectro-temporal representation of speech in primary auditory cortex. We represent a typical spike train in layer IV cortical neurons. Most of these neurons phase-lock to speech amplitude modulations. Response onset elicits a reset of theta oscillations in superficial layers (step 1) where auditory cortex output is generated. After reset, theta oscillations track the speech envelope (step 2). Theta reset induces a transient pause in gamma activity and a subsequent reset of gamma oscillations. Theta and gamma generators that are weakly coupled at rest become more strongly coupled and nested (step 3). Gamma power controls the excitability of neurons generating the feedforward signal from A1 to higher order areas (step 4). Neuronal excitability phase aligns to speech modulations (step 5): gamma tends to be strong when the energy in the signal is weak. *From Giraud & Poeppel, 2012*.

More recently, intracranial recordings allowed to specify which local cortical populations were involved in the STG and which characteristics of the speech signal envelope contribute most to the maintenance of the neuronal representation. The ECoG study by Oganian & Chang (2019) highlighted that the high gamma activity recorded in the

mSTG (see Figure 2B) responds preferentially to the discrete temporal cues of the peakRate, i.e. to the time points at which there is a maximum change in envelope amplitude (see Figure 5 below).

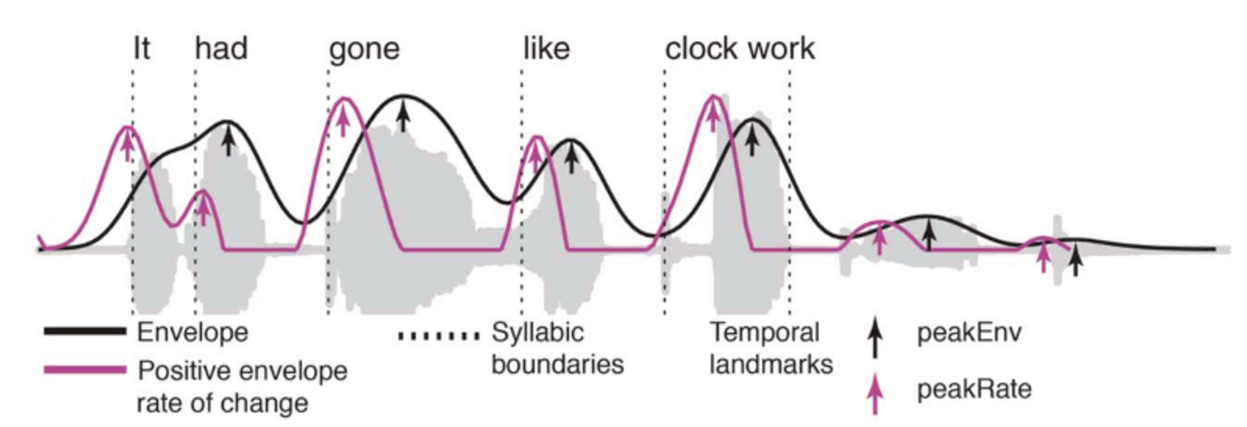


Figure 5 – Speech envelope and peakRate. Illustration of acoustic waveform of example sentence (from TIMIT stimuli database used for sentence-listening task), its amplitude envelope (black) and half-rectified rate of amplitude change (purple). Arrows mark local peaks in envelope (peakEnv) and rate of change of the envelope (peakRate), respectively. *Adapted from Oganian & Chang, 2019*.

Thus, neuronal responses in the STG reflect both the timing and magnitude of the peakRate events, thereby effectively encoding information about syllable structure, stress, and speech rate. These results suggest that at the scale of the local neuronal population, information on syllables (timing, stress and acoustic-phonetic content) can be considered distributed in several functionally distinct local neuronal populations, providing more nuance on the functional organization of the STG. Finally, a very recent study by Leonard et al. (2024) recording the activity of single neurons in the STG using a new recording technique (high-density Neuropixels arrays ; Jun et al., 2017 ; Chung et al., 2022 ; Paulk et al. , 2022) was able to confirm that all the observations concerning the processing of the different characteristics of the speech signal carried out on ECoG data (representing several thousand neurons) are also found at the level of the single neuron. Indeed, it has been demonstrated that we find selective neuronal responses specific to acoustic-phonetic characteristics (e.g. vowels or consonants) or to temporal acoustic cues such as onsets from silence or peakRate events.

Ultimately, it is clear that to approach a complete understanding of language it is necessary and essential to use language stimuli covering all its aspects and complexity. From the syllabic level through the words to arrive at the sentence level, passive listening tasks using these stimuli are fundamental and complementary to obtain a global and overall understanding.

c. Towards more naturalistic stimuli in speech perception studies

The degree of ecology of the stimuli that are used must be carefully considered if we want to be able to generalize the interpretations that are made in experimental setting, to the natural setting. Hamilton & Huth (2020) propose precise and very relevant criteria for estimating to what extent the language stimuli that we use are representative of stimuli that we could find in real life. The reasoning on the naturalness of a stimulus is based around three main questions: first, is it a stimulus to which a person could reasonably be exposed outside of an experimental setting? Second, does the stimulus appear in the same context as in real life? Third, is the participant's motivation to perceive and understand the stimulus specific to the experimental setting or is it a motivation that the participant would develop in real life?

Under this perspective, it is clear that stimuli like pure tones are artificial in all aspects, that syllables, although they are speech signals, are not stimuli that we are led to perceive in isolation in daily life. Isolated words are understandable and relevant per se, but they are mostly limited in real-life to imperative questions and statements. The presentation of isolated sentences represents a good degree of ecology when it comes to investigating the more complex processing of the speech signal, making it possible in particular to answer the question of the dynamic processing of speech, however, it is difficult to experience a particular motivation to understand decontextualised sentences and the context is strongly linked and essential for comprehension, especially when it comes to natural context. « Isn't it ? »

There are, however, more and more studies using stimuli pushing the cursor of the ecological framework to a level closer to that of real-life situations to study speech perception. The contextual aspect of speech is enhanced through the use of natural speech extracted, for example, from storytelling recorded during live conferences (e.g. The Moth: Simony et al., 2016 ; Huth et al., 2016 ; de Heer et al. al., 2017). This type of passive listening is in itself very representative of a real situation, the participant taking the same position as a spectator, the narration situations are quite common moments of everyday life, moreover the spontaneity of this type of extracts adds a little more to the immersion and motivation of the participant. More conventional stimuli are also used this time exposing participants to extracts from pre-recorded stories or extracts from movies (te Rietmolen et al., 2023 ; Hamilton et al., 2018). Other databases use synthetic voices in order to better control certain features of story-telling (Gwilliams et al., 2022 ; Gwilliams et al., 2023).

The multiplication of studies using natural speech in passive listening tasks clearly illustrates that there is a growing interest but that it's above all a necessity for the investigation at a higher level of speech processing, particularly the semantic level. However, a limitation still remains persistent, which is inherent to the passive listening paradigm, namely the lack of interactivity and the active role of the participant towards

someone else. These aspects will be discussed in more details in the second part (see The language in interaction section).

2. The production

The act of speaking is undoubtedly one of the most remarkable and distinctive features of the human species. Speech production is a complex and integrative behaviour, requiring the linking of high-level cognitive and linguistic processes to very fine motor control mechanisms. This has not always been the case. Knowledge of neural mechanisms involved in this language facet has considerably evolved since 1861, when Paul Broca first proposed that the foot of the third frontal convolution of the left hemisphere (now called Broca's area; see Figure 2A) was the center of spoken language. For a long time, the observation of deficits in speech production in patients with lesions in this area led to the belief that only this region was involved in speech production (and that, with a major logical flaw, this region only processed language). However, speech production encompasses a variety of processes that extend beyond this region.

Production starts with the intention to communicate, and continues with the translation of the message into lexical units (i.e. words) which are themselves decomposed into sequences of syllables ordered in time and adapted to the context in terms of rhythm, volume and prosody. The final stages require the fine coordination of several sensorimotor systems, including the respiratory system, which supplies the airflow needed to set the vocal folds into vibration, and the laryngeal and supra-laryngeal muscles, which convert this airflow into consonants and vowels by modulating the configuration of the vocal tract (articulation). Speech production is also the result of the interaction of sensorimotor mechanisms with linguistic processes, including grammatical, syntactic and semantic processes, as well as cognitive processes such as verbal memory and audiovisual attention (Tremblay et al., 2019). Speaking is thus an extremely rich and organised activity. In this section, a particular attention will be paid on the tasks developed to study speech production, the main models illustrating the various underlying processes and, especially, speech planning investigations.

a. Investigations and modelling of speech production

i. How speech production can be studied?

Several tasks have been progressively developed to investigate speech production, varying in their degree of complexity. The most widely used and documented is the picture naming task, which consists of naming orally (aloud or silently) a presented picture (Salmelin et al., 1994 ; Sörös et al., 2003 ; Christoffels et al., 2007 ; Costa et al., 2009 ; Volfart et al., 2022). Although not very representative of a natural context, this type of task involves lexical selection in order to find the word corresponding to the picture presented, as well as motor generation processes. In the same vein, there are also verb

and word generation tasks. In the former, subjects are asked to produce one or more appropriate verbs in reference to a noun (e.g. 'apple' - 'eat'), while in the latter a semantic category is presented to subjects, such as 'fruit', 'tools' or 'animals', and they are asked to produce as many examples as possible (Indefrey et al., 2000; Cappa et al., 2003; Indefrey & Levelt, 2004; Khader et al., 2010). As for picture naming, these tasks remain restricted in their degree of ecology but also involve all the essential components of isolated word production. Conversely, other tasks involving more than one word (similar to short simple sentences including adjectives such as 'the brown cat' or 'the big cat') can also be cited (Bürki & Laganaro, 2014 ; Riès et al., 2021). These tasks allow investigations of control and planning mechanisms of speech, the latter being more solicited than for isolated words since several words have to be produced in a short amount of time. Finally, more complex and specific tasks focus on the reading context and consist of reading relatively long words, sentences and texts (Indefrey & Levelt, 2004; Deniz et al., 2019; Bailey et al., 2021). In addition to investigating the various speech production processes, reading tasks - particularly those involving parts or entire texts - allow complex semantic aspects to be investigated in a more important way.

ii. Lemma model of lexcial selection and form encoding

Essentially based on investigations using picture naming tasks, Levelt et al (1999) developed and proposed a theory of speech production which has been and still is a partial reference point for this aspect of language. According to it, the architecture supporting word production is articulated in two main subsystems: the first for lexical selection, i.e. the identification of the most appropriate word at the level of lexical representation (mental lexicon), and the second for form encoding, i.e. the preparation of the articulatory form of the word.

The lexical selection subsystem comprises two initial processing stages. The *conceptual focusing* and *perspective-taking* stages involve converting the thought (speech intention) that we wish to express into the most appropriate lexical concept, i.e. into a semantic structure that constitutes the meaning of a particular word. With this comes *lemma selection*, which consists of mapping the selected lexical concept to the corresponding lemma, i.e. to an abstract word node that not only intervenes between semantics and phonology, but also points to relevant morphosyntactic features. The form encoding subsystem comprises three specific processing stages. Firstly, the retrieving morphemic phonological codes involves selecting and spelling out the segmental phonological representation of the word of interest. Secondly, *prosodification* and *syllabification*, which involves determining the metrical and syllabic structure of the word of interest, a progressive, context-dependent process that takes place 'on the fly'. Finally, *phonetic coding* and *articulation* involve the activation of pre-compiled units in the 'syllabary' and the generation of an 'articulatory score' to modulate the elements composing the vocal tract. With this, the model includes

a self-monitoring system consisting of two feedback loops facilitating the ability to detect and correct our own speech errors. A first 'external' feedback loop controls the auditory signals of self-produced speech and a second 'internal' loop controls the process of phonological word generation.

The important meta-analysis by Indefrey & Levelt (2004), composed by more than 80 neuroimaging studies based mainly on tasks involving the production of isolated words, such as picture naming, word generation (nouns, verbs) or word reading, highlighted the regions essentially involved and allowing to determine that it takes an average of 600 milliseconds to produce a word from the first lexical selection stage to the articulation (see Figure 6 below). However, there are some limitations to this model, notably the serial nature of processing levels. In addition, the self-monitoring process is not sufficiently developed, a point that is central in the DIVA model (Directions Into Velocities of Articulators ; Tourville & Guenther, 2011).

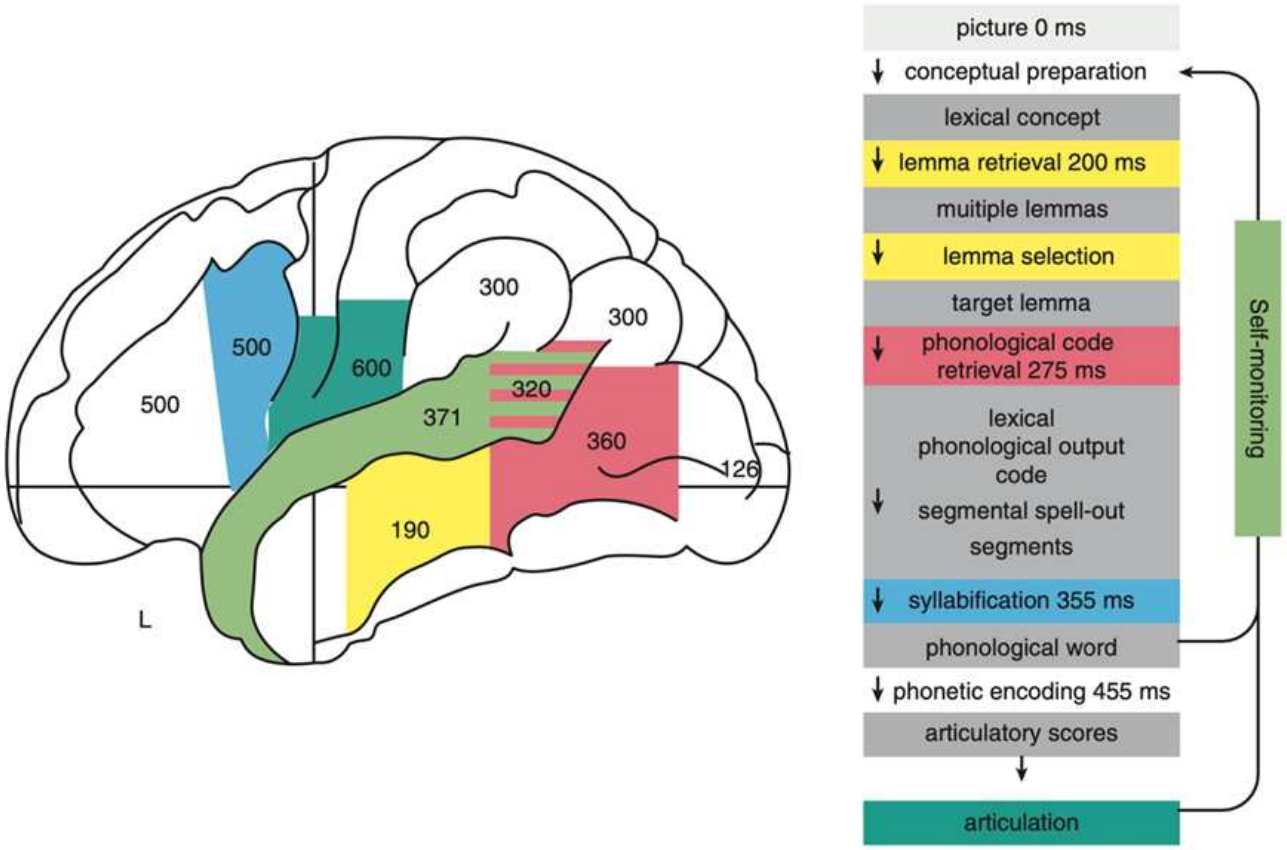


Figure 6 – Illustration of Indefrey & Levelt (2004) meta-analysis results. Left column: schematic representation of meta-analysis results for word production. Identical colours indicate relations between regions and functional processing components (right column). The numbers indicate the time windows (in milliseconds) during which the regions are activated in picture naming. Further regions involved in phonetic encoding and articulation are the right sensorimotor cortex, the right SMA, the left and medial right cerebellum, the left and right thalamus, and the right midbrain. A further region involved

in self-monitoring is the right mid superior temporal gyrus. Right column: time course of picture naming as estimated from chronometric data. *From Indefrey & Levelt (2004)*.

iii. Feedback control and predictive coding framework

In a complementary orientation to the reference model of Levelt et al (1999), the DIVA model (Directions Into Velocities of Articulators; Tourville & Guenther, 2011) proposes a more sophisticated definition of the phonetic encoding and articulation processes and gives an important place to the feedback mechanisms occurring during speech production. According to this theory, motor control of speech is based on two main subsystems: one corresponding to feedforward control, i.e. representing the activation of motor commands for articulatory gestures and their transmission to the vocal tract via the subcortical nuclei ; the other subsystem corresponding to feedback control, i.e. the use of auditory and somatosensory information from self-produced speech to detect errors and send corrective instructions to the articulatory component (see Figure 7 for more details about regions associated to each subsystem).

The feedforward subsystem is organised through the *Speech Sound Map*, which can be seen as a repository of acquired representations of speech sounds (essentially syllabic units) that serve as a starting point for articulation. Alongside this, the *Articulatory Velocity* and *Position Maps* contain representations of the vocal tract (those of the larynx, lip, jaw, tongue and palate) which enable the articulatory score of utterances to be specified. The *Initiation Map* represents a module which sends a "go" signal to the motor commands corresponding to the prepared speech. During speech production, activation of a particular unit in the *Speech Sound Map* engages the corresponding motor commands of the vocal tract in the *Articulatory* and *Positions Maps*, and these commands are also triggered by a 'go' signal from the *Initiation Map*.

The feedback subsystem is organised via the *Auditory Target Map*, a module containing representations of auditory targets (i.e. acoustic expectations) during speech production. Alongside this is the *Auditory State Map*, a module representing the auditory input linked to speech (including self-generated utterances). Linked to it, is the *Auditory Error Map*, which is a module that calculates the differences between the predicted sounds and the real sounds of auto-generated utterances. Finally, the *Feedback Control Map* is a module capable of adjusting or updating articulatory commands in the light of sensory feedback. During speech production, the *Speech Sound Map* not only sends instructions to the *Articulatory Velocity* and *Position Maps*, but also a prediction message to the *Auditory Target Map*, indicating how the utterance should ideally sound. The acoustic signals of the actual utterance are represented in the *Auditory State Map* and these signals are compared with the representation of the target (the prediction) by the *Auditory Error Map*. In this way, if the utterance has been produced correctly, the error map generates no output, but on the contrary, if the utterance has been produced incorrectly, the error map generates an error signal and alerts the *Feedback Control Map*,

which then sends corrective motor commands to the *Articulatory Velocity* and *Positions Map* to adjust the verbal production.

The somatosensory feedback circuit is organised in the same way as the auditory feedback circuit. The *Somatosensory Target Map* is a module for representing somatosensory targets (i.e. tactile and proprioceptive expectations) during speech production. The *Somatosensory State Map* is the module representing tactile and proprioceptive inputs during speech production. Finally, the *Somatosensory Error Map* is a module calculating the discrepancies between predicted and actual tactile and proprioceptive sensations associated with speech production. As in the case of hearing, the *Feedback Control Map* will operate in the same way, generating an error signal if the actual tactile and proprioceptive sensations do not match those predicted, and the motor commands will be updated.

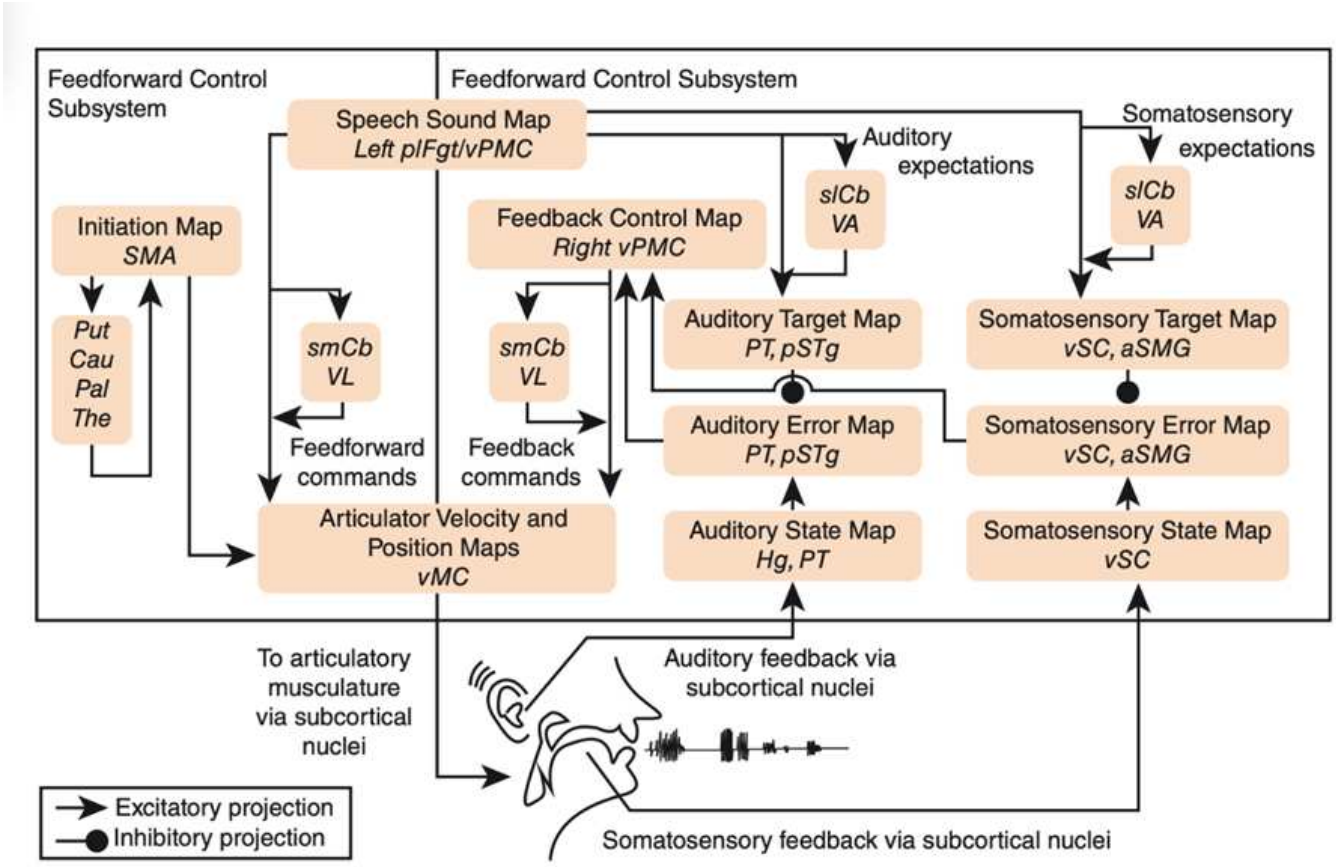


Figure 7 – Schematic of the DIVA Model of speech motor control. Each box corresponds to a set of neurons (or map), and arrows between the boxes correspond to synaptic projections that transform one type of neural representation into another. The model is divided into two basic subsystems: the Feedforward Control Subsystem on the left, and the Feedback Control Subsystem on the right. The neural substrates underlying this integrated control scheme include the premotor and primary motor cortices, auditory cortices, somatosensory cortices, the cerebellum, and the basal ganglia. pIFg = posterior inferior frontal gyrus; vPMC = ventral premotor cortex; SMA = supplementary

motor area; Put = putamen; Cau = caudate; Pal = pallidum ; Tha = thalamus; smCb = superior medial cerebellum; VL = ventral lateral nucleus of the thalamus; vMC = ventral motor cortex ; slCb = superior lateral cerebellum; VA = ventral anterior nucleus of the cerebellum; PT = planum temporale; pSTg = posterior superior temporal gyrus; Hg = Heschl's gyrus; vSC = ventral somatosensory cortex; aSMg = anterior supramarginal gyrus. *From Golfinopoulos et al., 2010, p. 863.*

The idea of feedback control processes that the DIVA model incorporates and develops is linked to a framework that is becoming increasingly popular and occupying an important place in the neuroscience of language, namely the concept of predictive coding (Friston, 2010; Heilbron & Chait, 2018). The core of this theory could be articulated around a question such as "How does the brain manage to make sense of the world?". Predictive coding sees the brain as a predictive structure that continually adapts to stimuli from its environment. Indeed, the brain constantly makes predictions about what exists, and improves them as it compares them with the real stimuli it perceives. Specifically, predictive coding proposes that the brain constructs a hierarchical and generative model of the world, a model capable of generating "top-down" patterns of activity that "bottom-up" external stimuli would trigger. In this way, the brain is constantly « active » in predicting sensory input. The comparison between the prediction and the actual stimulus generates a prediction error signal, which is larger when the the difference between the prediction and the stimulus is big. It is thanks to this error message that the predictions are then adjusted, and the estimates become more accurate in an error-minimization process. In this way, the generative model of the world is continually updated. Predictive coding therefore makes it possible to explain the way in which the brain processes sensory information and, in the context of language, allows to link the production and perception of speech. This is a point that has been investigated in particular through auditory feedback processes, where perception and production directly influence each other.

b. The case of auditory feedback

Auditory feedback is essential for speech production. In daily life, hearing one's own voice is a critical factor to maintain fluency in speech production (Tourville & Guenther, 2011; Yates, 1983) because this auditory feedback allows errors in vocalisation to be detected and corrected in real time. We continuously monitor our own voice and compensate for changes in auditory feedback, such as increasing the intensity and pitch of our voice in response to a noisy environment (Levelt, 1983; Lane & Tranel, 1971). Delayed auditory feedback tasks - in which the participant's speech signal is sent back to him/her in real time with a temporal delay and/or perturbations (e.g. fundamental frequency) - have revealed behavioural effects such as speech errors (repetition of syllables and words), modulations of speech rate, reading duration, adaptation of speech timing and increases in voice pitch and intensity (Lee., 1950; Black., 1951; Fairbanks., 1955; Stuart et al., 2002; Yamamoto & Kawabata, 2014; Karlin et al., 2021).

From a neurophysiological point of view, several observations relating to auditory feedback have been made. In particular, the phenomenon known as speaker-induced suppression (SIS), which is characterised by a reduction or even suppression of the neural responses (in the auditory temporal regions) corresponding to the vocal signal of one's own voice (or self-generated sounds) compared with the vocal signals of others (or external sounds) (Martikainen et al., 2005; Flinker et al., 2010; Niziolek et al. 2013; Kurteff et al., 2023). This phenomenon is in line with the description of the DIVA model and the predictive coding framework. Following on from this, delayed auditory feedback tasks have also shed light on complementary results. Neuroimaging studies using techniques such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI) have shown increases in neural responses in the superior temporal regions (pSTG) during delayed auditory feedback compared with normal feedback conditions (Hirano et al., 1997 ; Hashimoto & Sakai, 2003 ; Takaso et al., 2010). These increases in neural responses are interpreted as error signals related to the delay in auditory feedback that does not correspond to the motor commands used to generate articulatory movements. More recently, the study by Ozerk et al (2022) investigated the neural substrates of delayed auditory feedback in more detail using intracranial recordings (electrocorticography: ECoG). The task included different delay conditions (no delay, 50, 100, 200 milliseconds) and was organised into two reading sessions; a first session of word reading (3 syllables, e.g. "document"); a second session of sentence reading (8 words, e.g. "the cereal was fortified with vitamins and nutrients"). For both word and sentence reading sessions, a significant increase in neural response (high frequency activity) located in the superior temporal gyrus was observed, scaling with the duration of the feedback delay. These results can be interpreted as an error signal as much important the greater the time lag between the production of the speech signal and the auditory feedback.

In parallel, it is important to note that other structures involved in motor control play an important role in speech production and fluency maintenance, namely the basal ganglia and the cerebellum (see Silveri, 2021 for a review). More precisely, these structures, which are reciprocally interconnected, are involved in the forward model process that compares the expected and actual sensory consequences of an action, and thus in the unexpected sensory consequences of self-action (Ito et al., 2008 ; Schwartze & Kotz, 2013 ; Kotz et al., 2014 ; Kotz & Schmidt-Kassow, 2015 ; Johnson et al., 2019). It is therefore important not to consider the cortex as the sole seat of motor control (of speech) but also to consider deep structures such as the basal ganglia and the cerebellum, which play a particularly crucial role in motor correction processes.

By manipulating the auditory feedback of the speech signal, it is thus possible to investigate the neural correlates of speech control models, linking production and perception. In addition, by using recording techniques such as intracranial recordings, it is possible to explore more distinctly the regions involved and their dynamics in the processes of speech production and planning, a point that will be addressed in the following sections.

c. Focus on intracranial recordings

i. The role of sensorimotor cortex in speech production

The heart of production is above all linked to the concrete act of speaking. When we speak, we have to dynamically coordinate the movements of our jaw, tongue, lips and larynx. These are very fine motor operations, involving nearly a hundred muscles that rapidly shape and reshape our vocal tract to generate particular vocal patterns, producing successive speech segments to build words and sentences. The end result is fluid speech production (Bernstein, 1967 ; Fowler et al., 1980; Browman & Goldstein, 1989; Gracco & Löfqvist, 1994).

One specific region in particular has been shown to be involved in motor acts : the sensorimotor cortex. Pioneering studies by Foerster and Penfield (Foerster, 1931; Penfield & Boldrey, 1937; Penfield & Roberts, 1959) using electrical stimulation described the somatotopic organisation of representations of the face and mouth in the ventral sensorimotor cortex (vSMC). This region has thus been the subject of much interest with regard to the act of speaking, and more recently studies using intracranial recordings have investigated the involvement of this region in greater detail. To understand the functional organisation of vSMC in articulatory motor control, Bouchard et al. (2013) recorded neuronal activity directly from the cortical surface in patients implanted with high-density multi-electrode arrays in preparation for surgery related to their drug-resistant epilepsy. Using a simple task involving the reading of consonant-vowel syllables (e.g. /ba/, /da/, /ga/), the authors demonstrated that neuronal coding in the vSMC is linked to the presumed kinematics underlying speech sounds. As cortical representations are organised somatotopically, with individual sites tuned to a preferred articulator and co-modulated by other articulators, there is a similarity in the layout of articulator representations and that of the vocal tract. Thus, their results reveal the dynamic organisation of sensorimotor cortex during the generation of multi-articulator movements underlying our ability to speak. However, this study shows its limitations in the types of production carried out, i.e. simple syllables, which are not very representative of the language acts that we perform most in a natural context. To follow up on this, Chartier et al (2018) recorded high-density intracranial electrocorticography (ECoG) signals, once again in patients with drug-resistant epilepsy, by asking them to pronounce complete sentences aloud. By using complete sentences, a level of complexity is added, characterised by dynamics and coordination of articulatory movements that are not well understood when producing isolated syllables. In these conditions, the articulatory movements underlying the production of a speech segment can be influenced by previous and forthcoming speech segments, a phenomenon known as coarticulation. Thus, the articulatory movements attributed to a certain production unit will vary according to the (production) context in which that unit is. The aim of Chartier et al. was to demonstrate how neural activity in the vSMC represents the movements of the articulators of the vocal tract during continuous speech production.

To do this they estimated the articulatory kinematic trajectories (AKT, see Figure 8) specifically for each electrode in order to characterise the heterogeneity of themovements represented by the activity in the vSMC.

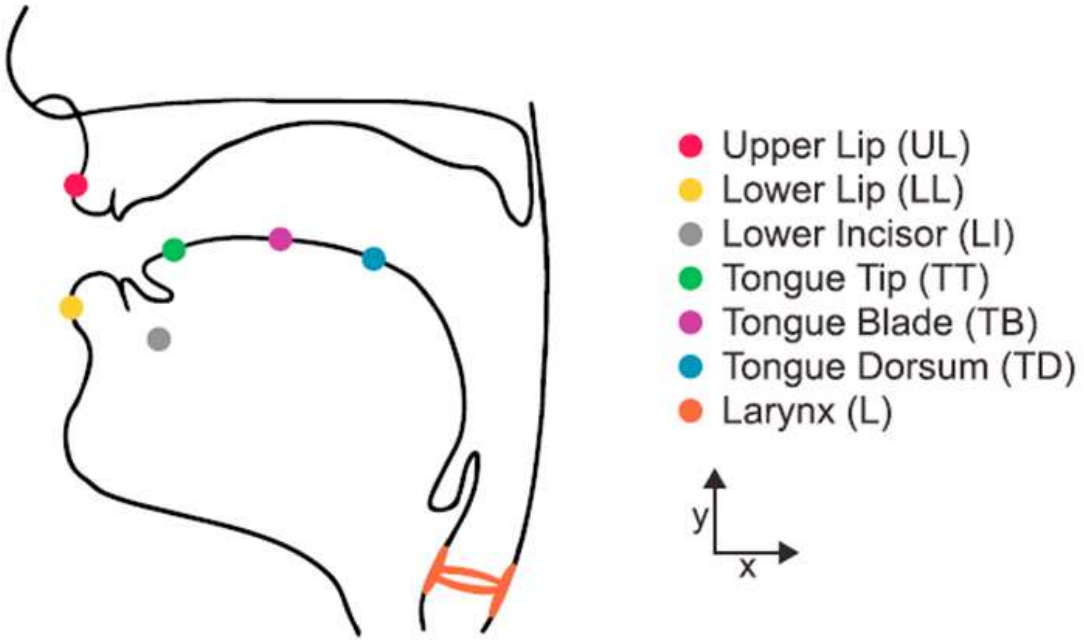


Figure 8 – Articulators used for Articulatory Kinematics Trajectories (AKTs). AKTs capture a wide range of movement types, but they can be differentiated by the location of vocal tract constriction. From Chartier et al. (2018).

Thus, by describing vSMC activity in relation to detailed articulatory movements it could be shown that distinct neural populations encode AKTs, representing a level of complexity that could not be observed using simpler, syllable-level speech tasks. Where the Bouchard et al. (2013) study examined vSMC activity on the basis of whether or not a given articulator was involved, Chartier et al. (2018) were able to show that, as with limb control, coordinated movements between articulators for specialised vocal tract configurations are also encoded in vSMC activity. These coordinated movements are at the origin of the complex kinematics underlying the continuous production of speech.

ii. Speech planning

Speech planning implies the processes operating just prior to actual production. Planning is a particularly event-rich process, encompassing diverse functions ranging from abstract conceptual and semantic operations to low-level articulatory programming and motor initiation (Indefrey, 2011). A large number of cortical regions have been associated with planning aspects, including the inferior frontal gyrus (Sahin et al., 2009 ; Schuhmann et al., 2009 ; Flinker et al., 2010), premotor cortex (Guenther, 2016), superior temporal gyrus (Basilakos et al., 2018), supplementary motor area (Alario et al., 2006 ; Guenther, 2016) and inferior parietal cortex (Mirman et al., 2019). Among the investigations performed by intracranial recordings, specific frontal regions have been particularly explored and tested to understand the processes of speech planning, notably the inferior frontal gyrus (IFG) or also historically called Broca's area. The ECoG study by Flinker et al (2015) highlighted that this region was not strictly speaking involved in the production of speech per se, contrary to the very first suggestion (Broca, 1861), but was rather involved in the processes preceding it, i.e. in planning. Using syllable repetition and word reading tasks (mono and bisyllabic) they showed that the IFG was active before and not during production, suggesting that Broca's area is not the seat of articulation per se but that it links representations of a spoken word to an articulatory code which is then implemented by the motor cortices responsible for coordinating the articulators (see previous section). In the same direction, using a focal cooling technique on the regions of the inferior frontal gyrus (pars triangularis/pars opercularis) and the precentral gyrus (motor cortex) in neurosurgical patients awake during a word production task, Long et al. (2016) showed a clear functional dissociation between these two regions. Focal cooling of the inferior frontal gyrus leads to changes in speech rate (speech timing) and the ventral part of the precentral gyrus leads to changes in vocalisation quality (articulation). These results once again suggest the role of the inferior frontal gyrus in the cascades of speech planning. Finally, more recently, the ECoG study by Castellucci et al (2022) investigated speech planning processes in more ecological contexts resembling interaction contexts. Using a question-answer paradigm (explained in more detail in section II.3.a *Turn-taking paradigms*) in which a single word (i.e. the 'critical information' or CI) initiates speech planning by providing the information necessary for a correct response, it was possible to specifically isolate the neural activities related to the perception, planning and production of speech with reference to turn-taking. Thus, Castellucci et al (2022) also showed that responses in the inferior frontal gyrus (in their case caudal IFG) occur mainly before speech initiation and that this activity is specifically associated with speech planning. In addition, they also demonstrated that the caudal region of the middle frontal gyrus (cMFG) is also involved in selective speech planning during interaction, thus providing new perspectives for investigations into these regions. All the results of this study, and in particular the functional and anatomical distinction between speech planning and perception – thus defining them as distinct modules – allow on a more theoretical level to suggest that the two processes can occur in parallel without interfering with each other. Consequently, as proposed in the review by Kuhlen &

Abdel Rahman (2023), it could explain how speech comprehension, planning and production can take place quasi-simultaneously in more dynamic and interactive contexts.

In conclusion, studies on speech production and planning mechanisms have progressively evolved, although restricted to isolated speech contexts. However, the use of increasingly complex and rich paradigms and improvements in recording techniques have led to significant advances, especially in speech production. In this instance, a recent study by Khanna et al (2024) using a new technique of ultrahigh-density Neuropixel recordings (Jun et al., 2017 ; Chung et al., 2022 ; Paulk et al., 2022) was able to sample activity across cortical columns in humans during production tasks. These latest advances herald new perspectives and even more precise and in-depth investigations into the regions of interest underlying speech production and planning.

II. The language in interaction

Language is a central topic in the field of neuroscience, and studies carried out in parallel on perception and production have made it possible to obtain – step by step and with progressively greater complexity – answers as to their underlying neural processes. However, the way in which language is commonly studied, i.e. in an isolated context, does not necessarily correspond to the way in which humans are exposed to and use it in daily life, and an essential aspect is still too often neglected, namely its use in an interaction context. Humans are social beings and we build ourselves through our interactions.

It is not very common to listen to white noise or pure sounds, nor to listen to simple syllables or words. It's not usual to produce words or sentences unless they are intended for someone else. The context is an important aspect and language is used primarily for communication. In this case, conversation has proved to be a perfect example for observing different phenomena of mutual adaptation between individuals.

1. Conversation as joint activity

a. Joint activity

A dance between two people, musicians in an orchestra, actors in a theater piece, a crew on a sailing boat - each of these situations requires a certain amount of coordination between the individuals to achieve the objective. It's important not to step on your dance partner's toes, to make the right movements at the right time, to adjust your movements according to what the other person is doing, to take into account your neighbour's next action, the next scene to be played out - in each of these situations it's important and necessary to be able to continually anticipate the behaviour of others and adjust our own. All interaction is a collective construction, and such situations can be associated with the concept of « *joint activity* » developed by Clark et al (1996).

In a joint activity there is an interdependence between the actions of the people involved, a complementarity in the sequences of actions to be carried out in order to succeed in the task. The words used by Klein et al (2004) perfectly describe what joint activity is based on : « Joint activity depends on *interpredictability* of the participants' attitudes and actions ». Whatever the type of task, success is reached when the parties involved are able to predict each other's actions. This interpredictability leads to another concept, namely *common ground*. *Common ground* (Clark & Brennan, 1991) represents all the shared information, knowledge, beliefs and assumptions between individuals that are relevant to the task of interest. At the beginning of the task, Clark et al (1996) define it as the *initial common ground*. Two professional dancers sharing a dance

will begin their dance knowing a priori that each knows the steps to be performed and the choreography. Comedians and actors will begin a piece in the knowledge that they all know the script. The *initial common ground* is therefore the common starting point, the expectations, presuppositions we have of each other when we start the activity, and the progress of the interaction leads to a gradual updating of this *common ground*, which can be called as *incremental common ground*.

Finally, coordination is a fundamental element in joint activity, directly echoing the anticipation of the other's actions and the adjustments to our own. This implies that there is a structure, an order in the succession of actions, stages to be respected, because part of the actions of one can affect the actions of the other. *Interpersonal coordination* thus represents a process of dynamic interdependence, of mutual adaptation, but also of adaptation of individuals to their environment. Based on these precise criteria, Keller et al (2014) have also proposed a model (see Figure 9) linking any type of rhythmic joint action requiring a degree of precision of the order of tens of milliseconds and temporal accommodation abilities. The case of ensemble music is a central example of their model - the case of dance could also very well be included - where severalfactors come into play during the performance of a joint action, such as cognitive and sensorimotor capacities like anticipation, adaptation and attention. However, these abilities would be affected by other factors such as the individuals' shared knowledge of the rules of the system (*initial common ground*), the degree of familiarity they share (e.g. the style of music or dance), the aim of the action and the strategy used to carry it out. Added to this, other very important factors are considered for the success of the interaction, namely social factors such as personality, social skills or empathy, which also affect the success of the interaction.

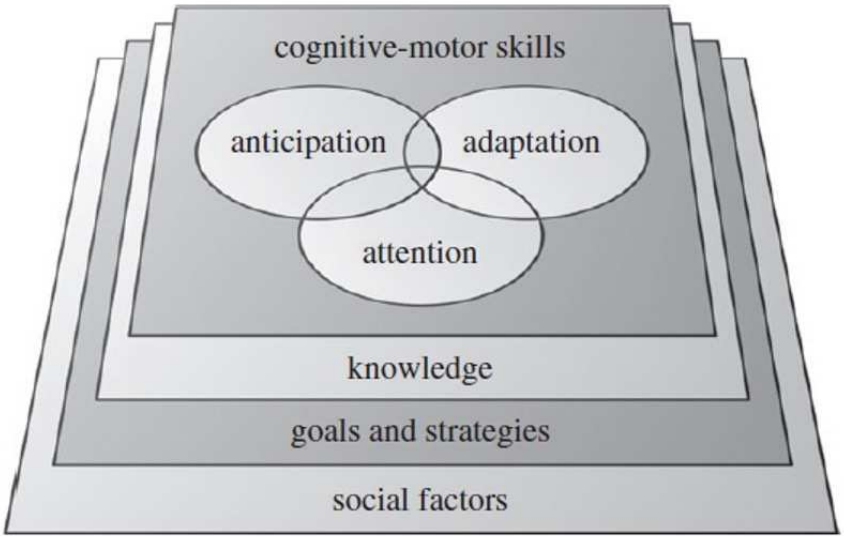


Figure 9 – Representation of factors affecting interpersonal coordination during a rhythmic joint action. *From Keller et al., 2016.*

In addition, Klein et al (2001) provide a relevant definition of the role of coordination in collective activities : "Coordination is the attempt by multiple entities to act in concert in order to achieve a common goal by carrying out a script they all understand". However, not all joint activities require an explicit common goal, and individuals are not even necessarily aware of whether a goal is present. Conversation is a perfect example of this ; it is also a rhythmic joint activity, generally undirected, its evolution depending on the topics discussed and the behaviour of each speaker, there is no exact predetermined scenario. In a way, conversation is a microcosm of coordination.

b. The case of conversation

Interestingly, although at first sight it is easier to describe the concept of « *joint activity* » through eloquent examples such as dance or music, Clark et al. (1996) developed it on the basis of conversation context. In fact, coordination aspects of conversation are found in its organisation into turn-takings, each interlocutor takes a turn to speak while the other listens and prepares a response. Sacks et al. (1974) place turn-taking as a major process in their model, defining it as a unit of social interaction constructed and reinvented by the interlocutors throughout the conversation. The turn can be defined in terms of two components: the first is defined by the verbal units, i.e. the words and sentences used by each speaker, these units can be highly variable in length from one turn to the next and are called *Turn Constructional Units* (TCUs). The second component is defined as the most relevant moment for the speaker transition, it is called the *Turn Relevant Place* (TRP) (Sacks et al., 1974). This model establishes several behavioural rules which must be mutually respected and understood by the interlocutors so that the turn to speak can take place at a time deemed appropriate by each party. It is on this last point that the coordination aspect takes on its full meaning, because the temporal organisation of turn-takings in conversation reveals a major fact: it is very rare for two people to speak at the same time or for several seconds to elapse between each turn. Interlocutors manage to coordinate themselves precisely so that one person's speech does not overlap with that of the other, or so that there is too much silence between turns. These principles are summarised by Sacks et al. (1974) as *minimal gap* and *minimal overlap*.

This means that the timing to speak in a dialogue is not random. The study by Stivers et al. (2009) highlighted and illustrated the temporal organisation of turn-takings across different languages and cultures. By studying spontaneous conversations in 10 typologically different languages, they selected sequences of closed response questions (i.e. yes-no) and subjectively measured, by asking a listening panel, whether or not the responses seemed delayed given the rhythm of the conversation. In addition, they objectively measured response times in milliseconds in relation to the end of the questions. These times are positive when there is silence between the question offsets and the onsets of the answers, and negative when there is an overlap of speech. By

analysing these temporal intervals across the 10 different languages, it was possible to show that a unimodal peak extending from 0 to 200 ms appeared systematically in each of the languages. In other words, although interlocutors belong to different cultures and speak different languages, they perceive a common temporal unit during their verbal exchanges, and the authors talk about a universal organisation of turn-taking. As a result, the ability of interlocutors to respect this common time between turns requires the ability to temporally anticipate and predict. The process of producing a word takes an average of 600 ms (Levinson & Torreira, 2015), and if the speakers manage to produce their speech within 200 ms of the end of the previous turn, it is likely that they have anticipated the end of the speaker's turn. In addition, when adult individuals are asked to determine - by pressing an answer button - the most probable moment for the end of several turns from spontaneous conversations, they estimate it on average 200 ms before the end of the turns. The *Floor transfer offset* (see Figure 10) is the name given to the distribution of the time elapsed between the end of the previous speaker's turn and the next speaker's turn to speak (Levinson & Torreira, 2015).

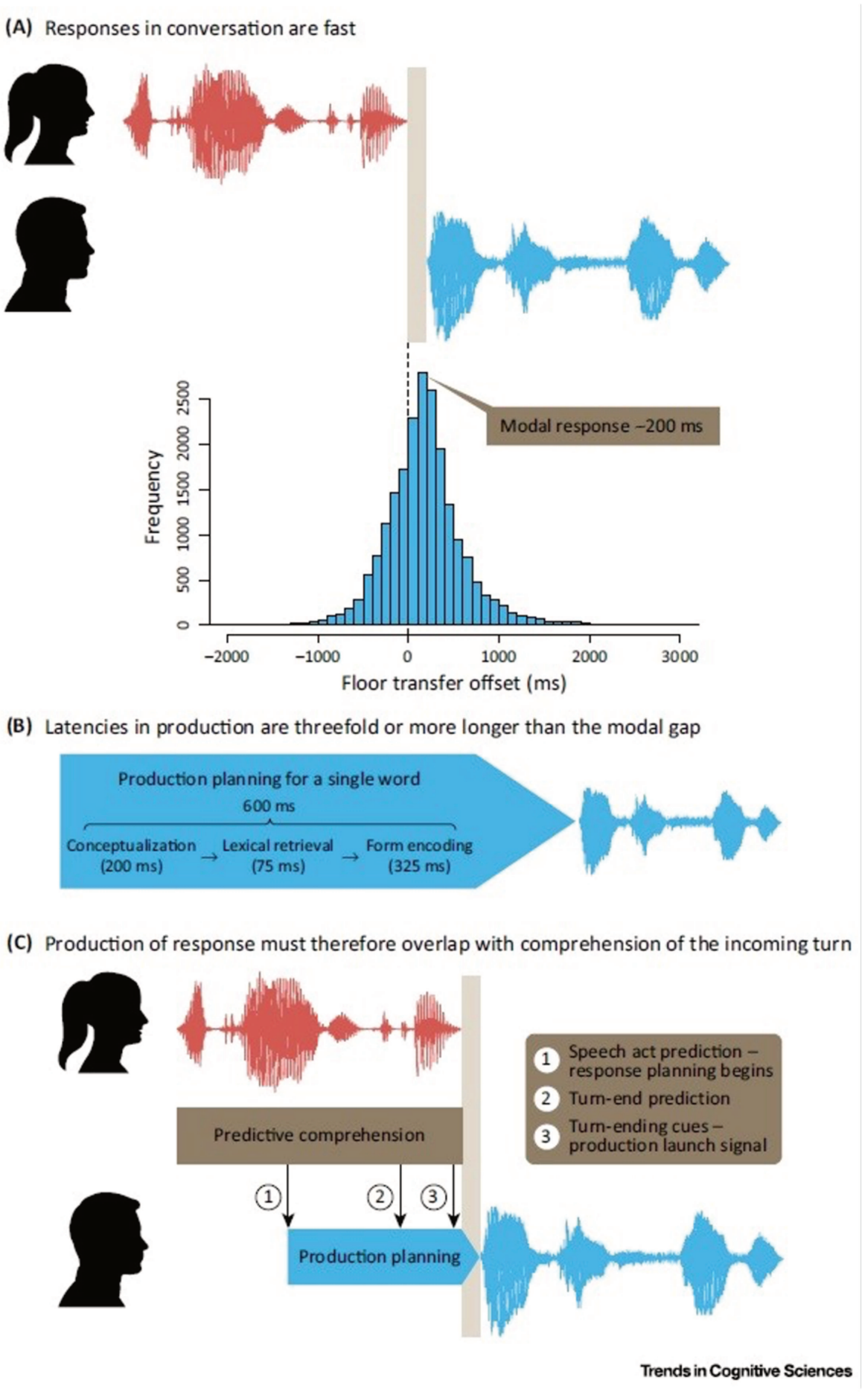

(A) Responses in conversation are fast
Modal response –200 ms
Frequency
0
500
1000
1500
2000
2500
–2000
–1000
0
1000
2000
3000
Floor transfer offset (ms)
(B) Latencies in production are threefold or more longer than the modal gap
Production planning for a single word
600 ms
Conceptualization (200 ms)
Lexical retrieval (75 ms)
Form encoding (325 ms)
(C) Production of response must therefore overlap with comprehension of the incoming turn
Predictive comprehension
1 Speech act prediction – response planning begins
2 Turn-end prediction
3 Turn-ending cues – production launch signal
Production planning
Trends in Cognitive Sciences

Figure 10 – Représentation du décours temporel des tours de parole. (A) Speech from the speaker A (in red) ends at point 0 of the Floor Transfer Offset (dotted line). The duration of the silences left by the speakers between turns is calculated from 0 to the start of speaker B's speech (in blue). The median inter-turn silence is around 200 ms. **(B)** Response latencies to produce single words, as measured in primed picture-naming tasks, require ~600 ms. **(C)**While speaker A is producing his message, speaker B is constantly generating predictions in terms of the content and timing of the message. Thus, in step 1 speaker B predicts the type of speech act (e.g. statement, question) that speaker A will produce according to the previous content. In step 2, the speaker predicts when the turn will end and in step 3, based on the acoustic cues at the end of the speaking turn, he determines when he will be able to start speaking. During these three different steps, speaker B can plan his response and produce it between 0 and 200 ms on average despite the planning-programming time required to produce a word. *From Levinson, 2016*.

Anticipating the end of turns is therefore essential, given the effort required during a verbal exchange. When one interlocutor speaks, the other must not only listen attentively to understand the content of the turn and prepare an appropriate response, but also maintain good timing so that the conversation flows smoothly. On the one hand, this coordination is co-constructed between the interlocutors, but it is also a multifactorial result. In addition to the content of the speaker's production per se, the roles of gaze, gestures, prosody and variations in the fundamental frequency will be essential, marking salient points on which the interlocutor can base his or her predictions. Cues from the interlocutor, known as feedback, are also important for the speaker to adjust his turn. They can be verbal (e.g. "yes", "i know") or vocal (e.g. mhm), but we can also mention smiles or head movements such as nods, which are very decisive. (Schegloff, E. A., 1982 ; Clark & Wilkes-Gibbs, 1986 ; Duncan, 1972 ; Ward & Tsukahara, 2000 ; Gravano & Hirschberg, 2011).

The context of conversation thus represents an event-rich place, giving way to phenomena of anticipation and prediction, requiring processes of production and comprehension. As with any joint action, conversation is a co-construction that takes place progressively between the interlocutors, and it is through their exchanges that very specific phenomena emerge that facilitate interaction.

c. Perception and production in conversation

The conversational situation represents a special case where perception and production are intertwined and take place over very short time windows, requiring the close integration of production and perception. Thus, the very fluid sequence of turns has put forward the argument that the representations underlying language comprehension and those underlying language production are similar. Pickering & Garrod (2013) have developed this point by placing the role of prediction at the center of their theory. To reconcile rapid and efficient comprehension and production, speakers must simultaneously predict their own action and that of their partner. Complementary behaviour occurs when co-actors (in this case interlocutors) use the same predictions to obtain different but coordinated behaviours. More explicitly, if actions need to be linked, one person can predict the other's behaviour by predicting what they themselves would do. Thus, the interlocutor must simultaneously predict the speaker's action (what he perceives) and predict his own action (what he must produce). This leads Pickering & Garrod (2013) to propose that during a verbal exchange, speakers use forward models to predict aspects of their next verbal productions and that listeners unconsciously imitate speakers and then use forward models based on their own potential verbal productions to predict what speakers are likely to say. In this way, this model helps to explain how it is possible to simultaneously reconcile the processes of comprehension and production, translating fluency in conversation.

So, as with any joint action, conversation is a co-construction that takes place progressively between the interlocutors, calling for prediction processes on both sides to ensure that the exchanges run smoothly. However, other very specific phenomena also come into play, fuelling and facilitating the mutual predictions of the interlocutors.

2. Convergence and interactive alignment

Interacting with someone else leads us to change some of our behaviours, and there is a tendency to imitate each other on different levels. Chartrand & Bargh (1999) have defined these reciprocal imitation phenomena as the "chameleon effect". This effect refers to the unconscious mimicry of the postures, mannerisms, facial expressions and other behaviours of one's interaction partners, so that one's behaviour passively and involuntarily changes to match that of other people in one's current social environment (Lakin et al., 2003). It has been shown, for example, that strangers who meet in order to carry out a joint task will automatically stamp their feet or rub their noses if their work partner displays one of these two behaviours (Chartrand & Bargh, 1999 ; Dijksterhuis & Bargh, 2001). Laughter, smiles, eyebrow movements and head nods are also more likely to occur if the other person has just used them (Louwerse et al., 2012). This type of imitation is called "*behavioural matching*" by Chartrand & Bargh (1999).

Thus, interlocutors in a verbal interaction influence each other through behavioural imitations of this type, but the conversational context will allow other – even more subtle – phenomena to emerge. In a context such as this, the general term "*convergence*" is often used, meaning that during their interaction, interlocutors will gradually see several of their linguistic representations converge. Moreover, in sociolinguistics, according to the Accommodation theory (Giles et al., 1991), this notion of "*convergence*" goes much further and defines the linguistic strategies by which interlocutors try to reduce their social differences. The words of Catherine Kerbrat-Orecchioni (1998) illustrate very well the phenomena that operate during conversation: "Speaking is exchanging, and it's to change by exchanging" (“Parler, c’est échanger, et c’est changer en échangeant”). There is a bit of the other in our words when we converse. The concept of "*interactive alignment*" proposed by Pickering & Garrod (2004) models this vision in detail, assuming that the conversation is successful to the extent that interlocutors manage to apprehend and understand the things they are talking about in the same way. More specifically, they construct mental models of the situation under discussion and a successful conversation occurs when these models align. The interlocutors do not generally align deliberately; rather, alignment is the result of the interlocutors' tendency to repeat each other's linguistic choices (see Figure 11).

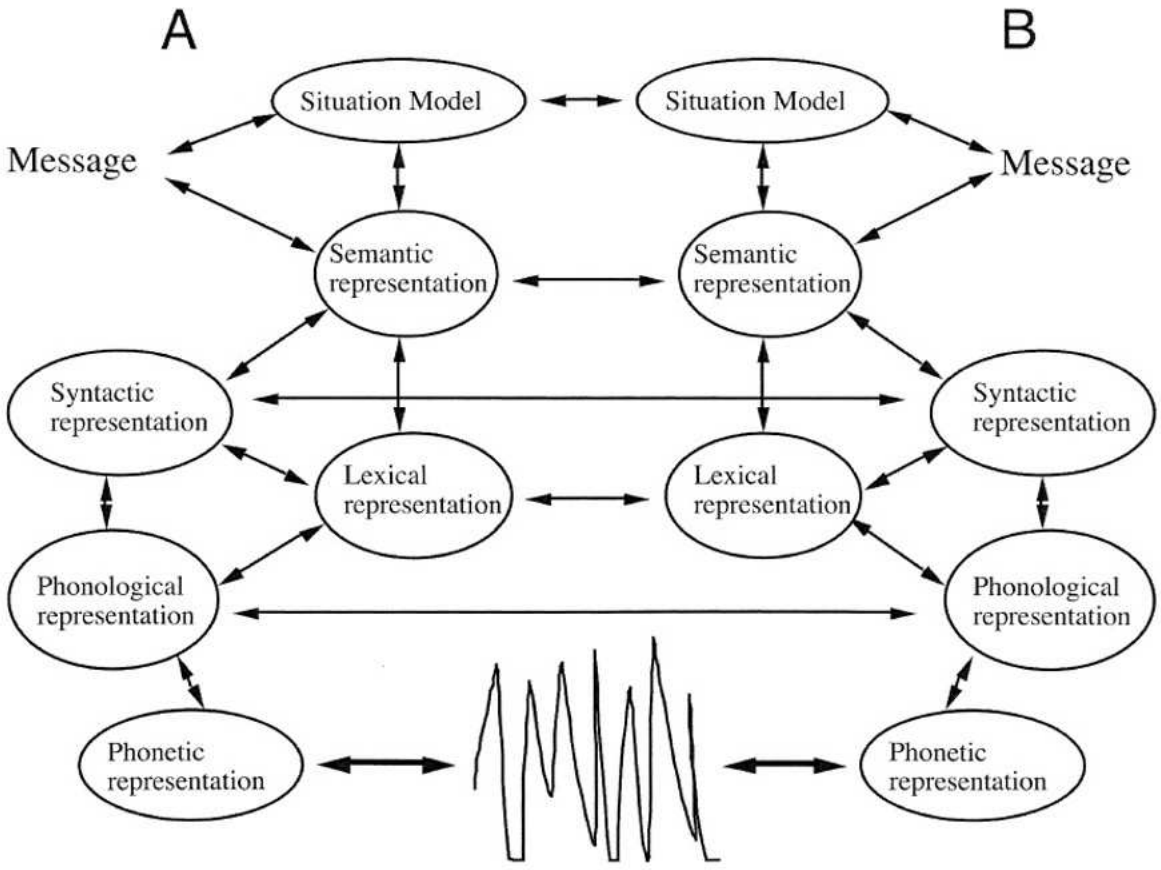


Figure 11 – Stages of comprehension and production processes in the Interactive alignment model. A and B represent two interlocutors in a dialogue. The schema illustrates the levels of linguistic representation computed by interlocutors and ways in which those representations are linked. There is an intra-locutors influence, between levels of representation, as well as an influence between locutors. Alignment occurs at a specific level when interlocutors have the same representation at that level. *From Pickering & Garrod, 2004*.

Pickering & Garrod (2004) propose that alignment emerges via a process of automatic, unconscious and effortless imitation at different linguistic levels. For example, when interlocutors describe images of objects to each other, they use the same expressions to refer to them (Brennan & Clark, 1996 ; Garrod & Anderson, 1987). This is, in itself, a very logical phenomenon: it is easier to understand what we are talking about when we each use the same words to define it, so the aligned use of a word is accompanied by a specific interpretation of that word. Markman & Makin (1998) concluded that people who describe things in the same way tend to think about them in the same way. Numerous other studies have also shown that interlocutors tend to develop the same sets of expressions to refer to particular objects, and that these expressions become more concise and similar when they are repeated with the same interlocutor and are notably modified when there is a change of interlocutor (Clark & Wilkes-Gibbs 1986; Wilkes-Gibbs & Clark 1992). Similarly, interlocutors tend to repeat the grammatical structures (syntactic level) chosen by the other to describe images of events (Branigan et al., 2000). If one interlocutor describes a card as "The chef giving the jug to the swimmer", the other will tend to describe the next card as "The cowboy handing the banana to the burglar"; but if one locutor says "The chef giving the swimmer the jug", the other will tend to say "The cowboy handing the burglar the banana". There is an alignment on the syntactic level which, in this type of exchange of descriptions, is very effective. If my partner describes an image in such a way, they are likely to understand me more easily if I describe mine in the same way.

In addition, Pickering and Garrod (2004) point out that alignment processes at different levels of linguistic representation (e.g. lexical, syntactic, semantic) interact in a way that increased alignment at one level leads to increased alignment at other levels. In such cases we speak of "*percolation*" between linguistic representations. For example, in the study by Branigan et al (2000), alignment is reinforced by word repetition, with speakers being even more likely to say "The cowboy handing the banana to the burglar" after hearing "The chef handing the jug to the swimmer" than after hearing "The chef giving the jug to the swimmer". This is a repetition of the word "handing". A simple example of percolation between the lexical and semantic levels is found in the use of a common word between interlocutors, leading to the understanding of an object, a person, a place, an event or an idea (conceptual alignment). It has also been found that grammatical alignment is strengthened for words that are semantically related. For example, Cleland & Pickering (2003) found that interlocutors were more likely to produce a sentence like "the sheep that's red" compared to "the red sheep" after hearing "the goat that's red" than

after hearing "the book that's red". Here, the words “sheep” and “goat” are strongly linked semantically, whereas the word “book” is much further away from them. In another example, the tendency to say "the sheep that's red" (rather than "the red sheep") is stronger after hearing "the goat that's red" rather than "the knife that's red". This shows that the semantic relations between lexical items influence the syntactic representation.

The different alignments at each level of linguistic representation and the percolation between them is what leads to an alignment of situation models between interlocutors. Added to this, the interlocutors imitate each other on non-linguistic levels, with tendencies to adopt the same postures (Shockley et al., 2003) but also to laugh and yawn together (Hatfiel et al., 1993) clearly meaning that interlocutors are also aligning on a non-linguistic level. Mutual understanding is therefore the ultimate goal of *interactive alignment*, making interaction between interlocutors more fluid and improving the effectiveness of the conversation. The use of language as we do in daily life is indeed an extremely rich activity, based on phenomena of mutual coordination that must be considered if we want to approach the language in its entirety.

3. Introducing interactivity into experimental paradigms

Previously, the interactive dimension of language has been discussed, highlighting in particular the importance of taking into account a number of phenomena such as anticipation, prediction and, more generally, the coordination that is necessary for the smooth running of a verbal interaction, specifically in conversation. However, these aspects are problematic when it comes to studying their neural correlates, the main constraint being the balance between the degree of ecology of the paradigm and the experimental control that must be provided. As a result, several paradigms have been set up to answer questions specific to language in the context of interaction, varying this balance according to the processes and phenomena of interest. The following sections will present different interactive paradigms grouped according to the interaction context and the recording techniques used.

a. Turn-taking paradigm

There are two major constraints when it comes to studying the neural correlates of language in a natural context: the technique for recording the brain signal and the experimental control of the task used. The intrinsically disordered and unpredictable nature of spontaneous conversations makes it very difficult to use standard methods of measuring brain signals such as electroencephalography (EEG) or magnetoencephalography (MEG). This is particularly true concerning evoked potential analyses (ERP), which generally require the repeated presentation of many similar stimuli so that the small variations in activity measured can be added up for all participants and all events in order to identify systematic fluctuations in a relatively small signal. The

structure of spontaneous conversation can't be restricted to this type of control. An additional difficulty arises during the production of speech, where the generation of muscular activity induces artefacts that can alter the quality of the signal. Unlike eye movement artefacts (e.g. blinks, saccades), which are very stereotyped and easy to detect and eliminate, this is not the case for electromyographic (EMG) artefacts, which can contaminate large parts of the EEG signal and are more difficult to identify and separate from brain activity (Vos et al., 2010 ; Jiang et al., 2019).

An intermediate option between the entirely natural context of conversation and the strictly isolated and controlled context has nevertheless proved to be a good compromise. As a verbal exchange is organised by turns, an interest developed in investigating these key moments. In fact, as it is characterised by an alternation between production and perception, the organisation in speech turns offers certain freedoms and possibilities for studying language in a less restricted and more ecological context. Thus, when speaking, three major cognitive processes can be counted: perception of the interlocutor's turn (speech comprehension), planning of one's own speech turn (speech planning) and finally production of our own turn (speech production). Since the process leading up to the production of a word requires at least 600 ms and the interval between speech turns lasts approximately 200 ms (Levinson, 2016), studies focusing on turn-takings are specifically interested in the neural correlates involved in anticipating the end of turns and in the planification of the next one. Among these investigations, several paradigms have been set up, in particular those aimed at simulating verbal exchanges, turn-taking situations and creating an interactive context.

A first example is the EEG study by Boux et al (2021), where the aim was to study brain activity prior to the production of different types of speech acts (language game task), in this case the actions of requesting and naming objects. The objective was to make a distinction between an isolated speech act, outside an interaction context (object naming) and a speech act directed towards someone else, for the purpose of communication (object request). To this end, they developed an object request and naming task performed by uttering single words integrated into role-playing with a partner, which resembles natural communication. In the naming condition, the participant had to imagine that they were taking part in a language test by interacting with an examiner, who assessed whether the real objects placed on the table were correctly named by the candidate. In the request condition, the participant had to imagine that they were interacting with a salesperson and that, as a customer, they had to request items for purchase. In both conditions, the participant and the confederate were seated on either side of a table where two objects were placed. After an auditory signal, the participant had to mentally select one of the two objects and then fix his gaze on a red dot located at the center of the table in order to restore neurophysiological activity to a baseline. After an interval of a few seconds (of the participant's choice) and still gazing at the red dot, the participant had to either ask for or name the object by saying just one word, for example "bottle" (see Figure 12). In this way, Boux et al (2021) looked at EEG activity 2000 milliseconds before and 500 milliseconds after voice onset and were able to show that around 600 milliseconds before the start of speech, an evoked potential

located in the fronto-central motor regions was greater during the preparation of requests rather than naming actions. They interpreted this as indicating the communicative intention of the speaker, i.e. the communicative function of the planned speech act, reflecting the speaker's expectation of the partner's actions that generally follow requests, for example the gesture of handing over a requested object. The example in this study thus shows the importance of taking the context into account, the only difference between the request and the naming being the intention of the speech act and the situation (role-play) in which this act is performed. This opens up possibilities for analyzing the neural substrates involved in situations that are both more ecological and controlled.

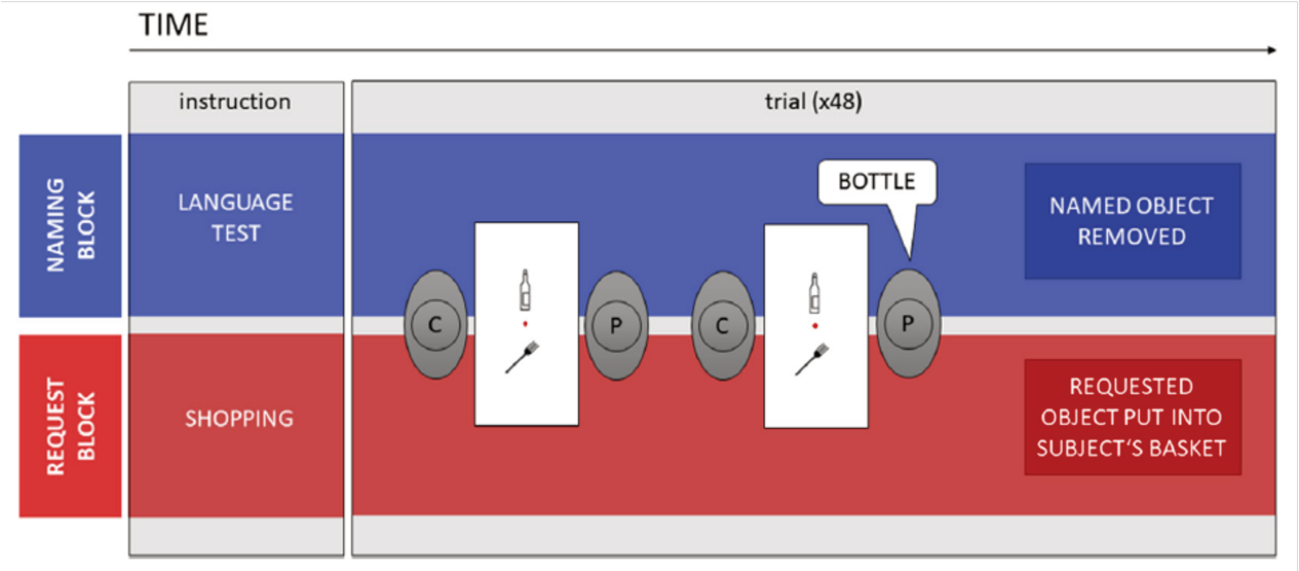


Figure 12 – Schematic representation of the structure of experimental blocks in the naming and request conditions. *From Boux et al., 2021*.

Other works have also proposed ingenious paradigms that increase the ecological context slider. For example, in the study of Bögels et al (2015) the authors designed an interactive turn-taking task in the form of a quiz game. Participants had to answer general knowledge questions using a single word or very short sentences, a classic and very common situation in conversations. In order to investigate the neural correlates underlying the response planning process, the participants' EEG signals were recorded. Two types of question were pre-recorded by the experimenter for two respective conditions: 1) early: the key information for answering the question was positioned in the middle of the question, for example: "Which character, also called 007, appears in the famous movies?"; 2) late: the key information only appeared at the very end of the question, for example: "Which character from the famous movies, is also called 007?"; the key information being the word "007". In this type of paradigm, the aim is to investigate the period before the answer is given, in this case by looking at the time at which

participants begin to plan the production of their answer by varying the position of the key information needed to find and produce the answer. The interactivity of the task lies in the presence of another person : the experimenter asks the questions and the participants answer. Some of the filler questions (but none of the experimental questions) were recorded with a hesitation or an error to create the illusion of live speech. In this situation the use of language is directed towards a communicative goal, it is not a context of isolated production, there is a concrete objective and a level of control is retained. They studied the temporal evolution of production planning, and the main results supported arguments in favour of early planning of turn preparation (response). The main results centered on a significant positive evoked potential (ERP) and a decrease in power in the alpha band (9-14 Hz) as soon as the information crucial to answering the questions was available (see Figure 13). Bögels and collaborators (2015) interpreted these effects as reflecting a shift in attention from understanding the question to planning the production of the answer, concluding that the planning of the next turn of speech takes place as soon as possible (in reference to early condition).

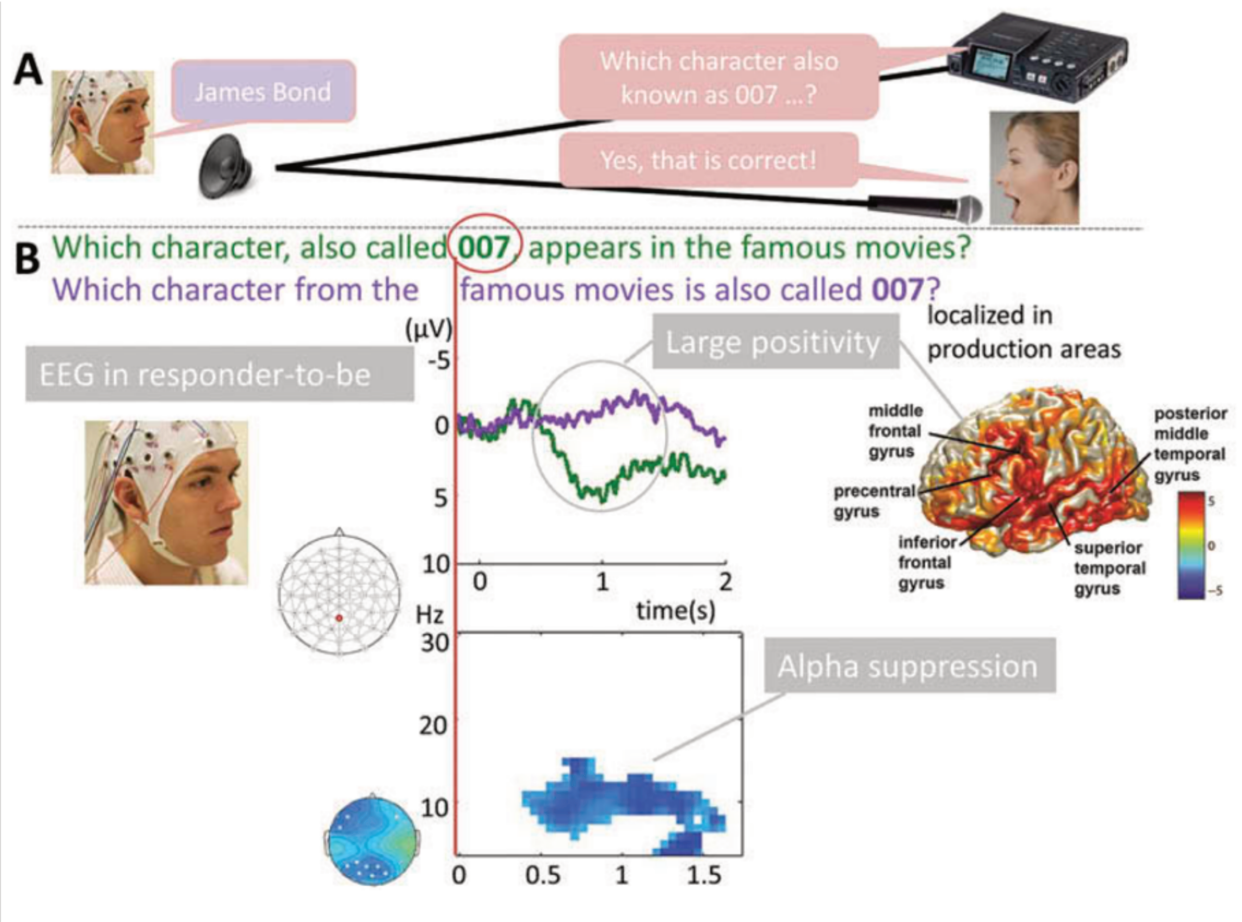


Figure 13 – Main results of Bögels et al. (2015). Setup (panel A) and results (panel B) for the ERP (top) and time-frequency domain (bottom). *From Bögels et al. 2017* (see also Bögels et al. 2018 for a replication of the results).

Another EEG study by Bögels et al. (2020) subsequently aimed to replicate the results of the previous studies (Bögels et al., 2015; Bögels et al., 2018) but this time in a more natural situation. In the previous paradigms they carefully manipulated the turns in order to artificially create a distinction between 'early' and 'late' planning, here they adapted the paradigm by mixing scripted and unscripted questions while keeping this trick inciting early or relatively late response planning. The procedure was very well designed: first, the participants were asked to fill in a short questionnaire containing some personal information to personalize the main questions in the task. Next, the experimenter was asked to ask eight scripted questions which could appear in two different versions, a long and a short version. The participants heard four short questions and four long questions, the latter being chosen on the basis of the answers to the questionnaire. For example: "So you are a student?" (short); "So you are a student here at Radboud University?" (long) or "So you play volleyball?" (short); "So you play volleyball on Thursday night?" (long). These questions led to closed answers (yes-no). After each scripted question, the experimenter freely asked a number of unscripted questions related to the subject addressed by the scripted question (for example, asking what kind of study they were doing, who they played volleyball with, whether they were competitive, etc.). Once the subject was closed, the experimenter moved on to the next scripted question and the procedure was repeated. In this paradigm the aim is to introduce as much spontaneity as possible, and mixing scripted and unscripted questions reinforces this impression of natural conversation. Thus, this study provided further support for the early planning theory of turn preparation, replicating the result found in the study by Bögels et al. (2015 ; 2018) regarding the broad positivity observed in favour of questions prompting early planning. This provides further support that it is possible to use a context closer to a spontaneous dialogue to answer the research questions while retaining a certain level of control.

Finally, another original type of paradigm was also used in a study by Fjaellingsdal et al. (2020) where participants had to construct a four-word sentence in German with a partner, taking turns for each word in the sentence. Called the word-by-word paradigm, it is inspired by improvisational theatre exercises in which two speakers jointly produce a sentence. This paradigm therefore introduces an interactive context simulating a verbal exchange between two people and enables behavioural aspects to be assessed, such as speech times and electrophysiological responses such as event-related-potentials (ERPs). In the case of this study, the procedure was as follows (see Figure 14): the participant and partner were seated in the same room, each in front of a screen, the whole sentence was displayed for the partner and he had to pronounce only the first and third words, while the participant had to pronounce only the second and fourth words (a word displayed on the screen during the appropriate turn with its associated image). In 30% of the trials, the partner pronounced an article of unexpected gender and to construct a correct sentence the participant had to adapt to the unexpected article by producing the synonym corresponding to the article (see Fjaellingsdal et al. 2020 for more details). The results will not be reported here, but this study also succeeded in showing

that it is possible to combine electrophysiological measurements (here ERPs) of verbal interactions while maintaining sufficient experimental control.

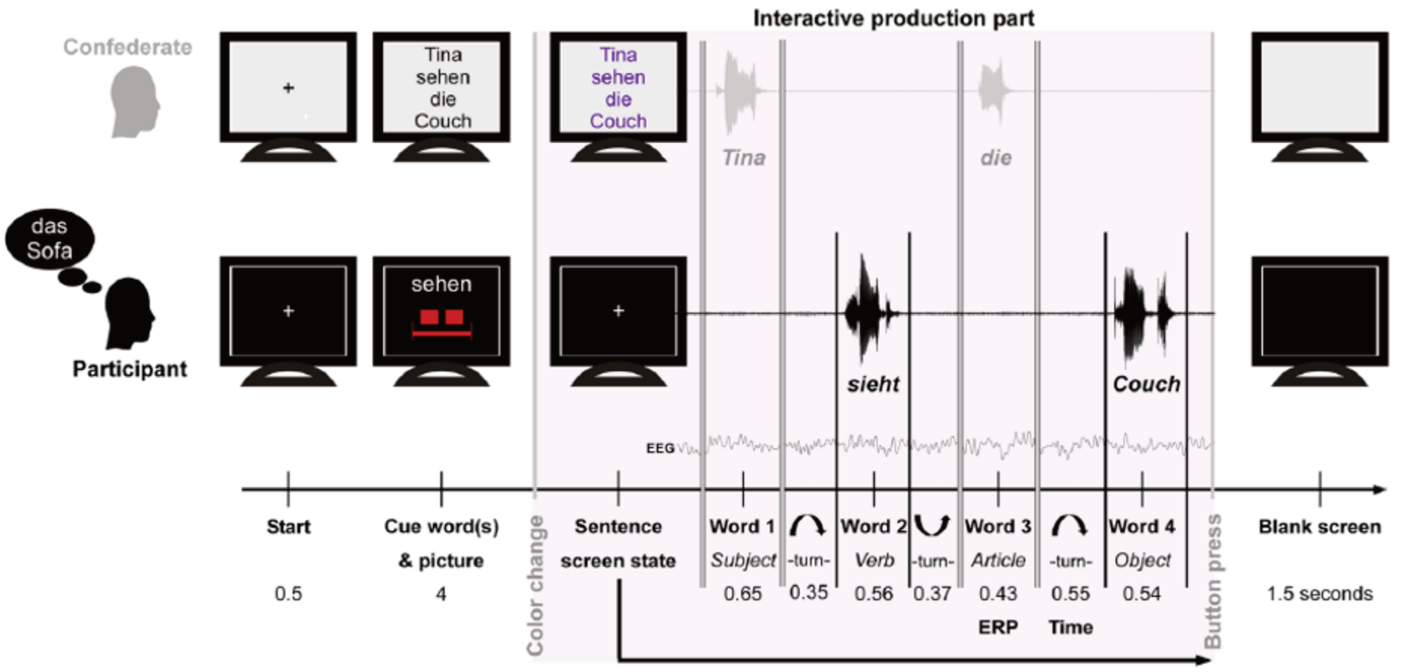


Figure 14 – Word-by-word paradigm timeline. Screen states and streams of participant and confederate. Example of an unexpected trial: participant named the picture '*das Sofa*', expects the article '*das*' and hears the article '*die*'. Steady presentation times of fixation cross and words and picture are given in seconds. During the interactive production part (lilac box), the participant saw a steady white fixation cross. The confederate started the sentence with word 1 (sentence subject), the participant then uttered word 2 (sentence verb), the confederate continued with word 3 (sentence article) and the participant ended the sentence with word 4 (sentence object). Average word duration for each word of a sentence, as well as turn-times from word to word are given in seconds. The ERP is computed for the expected/unexpected article (i.e., time-locked to the onset of word 3). The response time is contrasted between offset of word 3 until onset of word 4 after expected and unexpected articles. Sentence translation: "*Tina sees the couch*". *Adapted from Fjaellingsdal et al., 2020.*

Taken together, these different paradigms illustrate how it is possible to introduce interactivity into language tasks while retaining a degree of control. This allows compatibility with brain signal recording techniques such as electroencephalography by minimising artefacts caused by muscular activity generated during speech production and thus enabling analysis by evoked potentials (ERPs). Although this does not represent a fully natural context representative of spontaneous interactions, the turn-taking structure facilitates this and provides particularly ideal periods for investigating the neural processes involved in the motor planning of speech production.

b. The case of hyperscanning

Previously, paradigms focusing on the interactive and/or conversational context have been approached, but in each case the aim was to investigate the neural correlates of a single person taking part in an interactive activity. Research into social interactions in general and the interest in investigating common activity specific to the social context thus motivated and made compulsory the development of a new approach, called 'hyperscanning'.

i. Principle and method

The simultaneous measurement of the haemodynamic or neuro-electrical activity of two individuals was introduced and applied by Montague et al (2002). Referred to as 'hyperscanning', this method aims to investigate the neural correlates involved when individuals interact in social activities (social cognition) and particularly to study the similarities between individuals' brain activities. The first hyperscanning studies focused on situations in which individuals engaged in imitation, cooperation and competition tasks (Montague et al., 2002 ; Dumas et al., 2010). In order to study these correlates underlying social interactions, it is necessary to use experimental paradigms that involve individuals in realistic, structured social interactions that are ideally ecologically valid and take place dynamically. This last aspect is central to hyperscanning studies. This fast-growing field in neuroscience covers a wide range of topics and answers questions that extend well beyond language, such as decision-making, problem-solving, knowledge-sharing, joint attention, affective communication, emotional regulation and musical practice (Liu et al., 2018 ; Redcay & Schilbach, 2019 ; Czeszumski et al., 2020 ; Kelsen et al., 2022).

A variety of methods are used to record brain signals in hyperscanning studies, including electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS). The latter technique uses optical sensors that measure non-ionising light in the near infrared at different points on the surface of the scalp to measure variations in cerebral blood flow, an indicator of blood oxygenation levels similar to fMRI. Significant advantages of fNIRS are its mobility and its signal, which is not strongly influenced by the participants' movements. These criteria make it easy to set up experimental paradigms that more closely resemble real-life situations. Of the many studies carried out in hyperscanning, the majority use fNIRS, and EEG is also well represented (Czeszumski et al., 2020 ; Kelsen et al., 2022 ; TenHouten et al., 2023 ; Hakim et al., 2023 ; Grasso-Cladera et al., 2024), as these techniques do not constrain participants in a scanner as is the case for fMRI and MEG.

Hyperscanning research focuses on social interactions in dynamic situations, so all periods of interaction are investigated. Because of the nature of the experimental paradigms used, there is no repetition of the same stimulus, so analyses of evoked potentials (ERPs) are not possible. In this field of study we are more interested in the evolution of the neural dynamics of the participants during their interaction. Specifically, the aim is to evaluate the co-variation of neural dynamics between brains of the participants, where these dynamics are analyzed by frequency bands (delta, theta, alpha, beta, gamma). Different terms such as 'brain-to-brain coupling', 'interpersonal neural synchrony (INS)' or 'interbrain synchrony (IBS)' are used in hyperscanning, but they all relate to defining the same principle, which refers to the patterns of brain activity that appear in two (or more) individuals during their social interactions. The purpose of the analyses carried out is to quantify the degree of similarity and synchronization between brain signals, and to quantify the correlation between these signals as a function of the different brain regions. To do this, the main analysis methods used are phase-locking value (PLV) and wavelet transform coherence (WTC). Other analyses are also used, such as functional connectivity, information transfer and mutual information analyses using, for example, granger causality (Czeszumski et al., 2020 ; TenHouten et al., 2023 ; Hakim et al., 2023 ; Grasso-Cladera et al., 2024).

As the research topics are extremely broad in hyperscanning studies, in the context of this thesis introduction we will only focus on examples of studies where the research question is directly related to the use of language in an interaction context, and where the paradigms put individuals in a direct communication situation.

ii. The use of hyperscanning to study language in interaction

One of the particularities of hyperscanning studies is that as the main objective is to study the cognition of several individuals in a social context, the paradigms that are set up must be as ecological and similar to real-life conditions as possible. This necessarily leads to a lesser degree of control, although still present, and the dynamic aspect of interactions is investigated in a more global and therefore less discrete way than in more finely controlled studies (see section 3.a).

A first example of a study illustrating key aspects of hyperscanning is that conducted by Pérez et al (2017), in which participants' brain activity was recorded using EEG. The task was simple, participants were asked to exchange their opinions and preferences regarding five different common topics: sport, cinema, animals, music and travel. The paradigm chosen simulated a particular conversational situation. It was a scenario similar to a telephone conversation in which the participants, seated on chairs side by side, could hear each other but were separated by a board preventing them from seeing each other in order to avoid possible non-verbal influences such as glances, gestures, postures and also lip movements (see Figure 15). The total duration for each topic of conversation lasted 180 seconds, with one participant taking on the role of speaker in the

first half and expressing what he or she had to say, while the other took on the role of listener and had to listen very attentively. They were asked to speak as freely and naturally as possible on the topic of conversation with the intention of addressing it to their partner. To help structure and guide their speech, a series of six questions were displayed on a screen in front of them. These same questions were also used at the end of the experiment to assess their degree of attention to their partner's stories. The EEG signals were analyzed in four frequency bands: delta (1-3 Hz), theta (4-8 Hz), alpha (9-12 Hz) and beta (15-20 Hz). The brain-to-brain coupling was calculated for each conversation topic between the speaker/listener pairs and for all EEG signal combinations (in their case 27 x 27 electrodes) using the phase-locking value (PLV). In hyperscanning there are several ways to assess the statistical significance of the degree of coupling or covariation of neuronal activities, the most commonly used method being to compare real dataset with a surrogate one. In this study the surrogate data was created by associating signals from different trials but conserving the relation between the pairs of participants as well as the role of each, i.e. for the pair of participants A/B, the EEG signal from participant A listening to talk about 'sports' was associated with the EEG signal from participant B talking about 'movies'. Here the relation between participants in the same dyad is preserved ; another possibility would have been to associate EEG signals between participants in different dyads, an option that is used in other hyperscanning studies. The coupling between the EEG signals and the amplitude envelopes of the speech signals was also calculated using PLV and a comparison with a surrogate dataset. Thus, as before, the amplitude envelopes were associated with EEG signals from different trials. A key point in all hyperscanning studies of speech production and perception (and for all forms of audio stimuli) is to distinguish brain-to-brain coupling that is not explained by the perceived auditory stimulus. So here, the electrodes significantly involved in brain-audio envelope coupling have been entirely excluded from the brain-to-brain coupling results. In this way, only the brain-to-brain coupling between the speaker and the listener, which is independent of the physical characteristics of the auditory stimulus, is interpreted. However, it’s important to take into that in EEG all channels are strongly correlated, and we may wonder whether this is sufficient. Moreover, as these characteristics are strongly found in the low frequencies (Ding et al., 2016), no brain-to-brain coupling effect is generally expected in the delta and theta bands, which makes it very difficult to investigate any significant brain-to-brain coupling effect within these frequency bands. Regarding the main results of this study, the authors emphasized on the one hand that there is indeed a coupling in the low frequencies of the brain signal with the speech signal (between speaker and listener) and on the other hand highlighted that part of the synchronization patterns found in the alpha and beta bands seems to emerge directly from the mutual interaction of the dyads without being influenced by the physical properties of the speech signal. Like the vast majority of hyperscanning studies focusing on language in social contexts (Kelsen et al., 2022), the evidence for the emergence of brain-to-brain coupling caused (in part) by mutual interaction ties in with theories and models postulating that brains function differently in interactive social situations than in a context of isolation (Hasson et al., 2012, 2016 ; De Jaegher et al., 2016 ; Redcay & Schilbach, 2019).

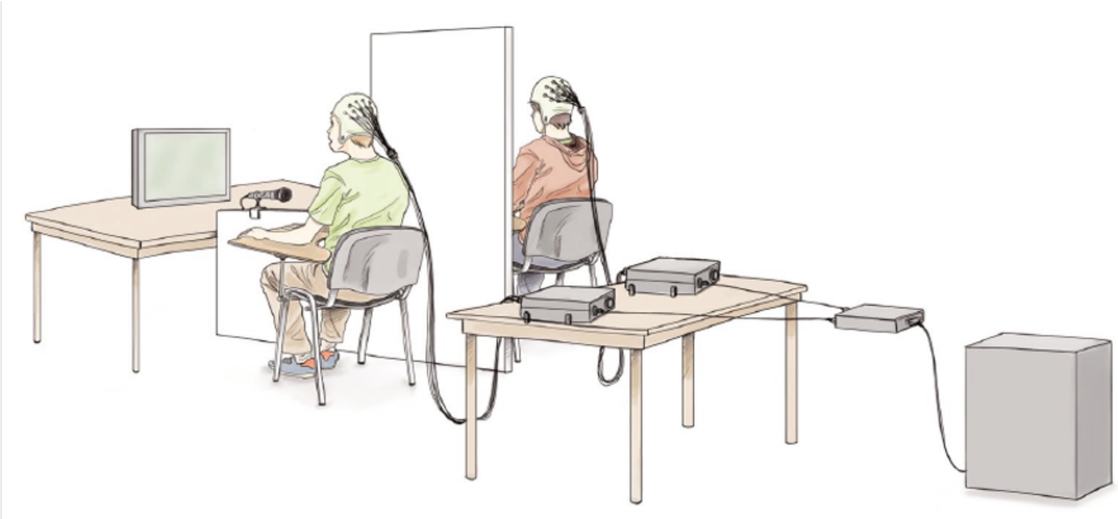


Figure 15 – Setup of the dual-EEG paradigm from Pérez et al. (2017).

A second example is the dual-EEG study of Mukherjee et al. (2018) where they were interested in the phenomena of implicit convergence taking place during conversation situations. They thus developed a relatively restrictive interactive and collaborative task called verbal domino task (VDT). During it, two participants are alternately presented with two words on a screen and must read aloud the word whose initial syllable coincides with the final syllable of the word previously produced by the other participant. It is therefore a question of chaining words in turn according to a rule of phonetic rhymes, precisely here it was a sequence of disyllabic words where there was a correspondence between the last syllable of each word and the first syllable of the following word in its spoken form, written form, or both. For example, after hearing participant A pronounce the word “delay”, participant B is asked to choose between the words “layout” and “tissue” and must pronounce “layout” (see Figure 16). The task is called collaborative to the extent that the participants must make the right choices to achieve a common objective, namely building the entire domino chain together and it is relevant for observing convergence phenomena due to the common syllables on pairs of words, influencing the participants' productions in a subtle and unconscious way. This type of paradigm was chosen because it shares certain characteristics of conversational interactions, although it fundamentally focuses on the phonetic aspect. Indeed, it allows to have control over the linguistic material used by the participants in order to more easily study convergence phenomena (see Convergence and interactive alignment section). Here the objective was to relate (phonetic) convergence phenomena with neural correlates in an interactive and dynamic task. To do this, the experiment was divided into three sections, the verbal productions of the participants were recorded individually before and after the duo session of the VDT task. The solo session speech data was necessary to establish a baseline for each participant. Thus, the solo task asked participants to read aloud 40 words (presented on

a screen) randomly chosen from the 200 words in the domino chain. In the duo session the task began with the display of a word on a screen of one of the two participants, while the screen of the other one remained empty. The first participant read the word aloud, his screen became blank, two words appeared on the screen of the second participant who had to choose the correct one according to the rhyme rule and this cycle was repeated until the end of the session. Thus, it was possible to compare the acoustic properties of the verbal productions of each participant, as the task progressed. These characteristics corresponded in this study to the mel frequency ceptral coefficients (MFCCs) which are a short-term representation of the power spectrum of sounds. For a pair of words to be convergent, the acoustic properties of the words for both speakers must become more similar to each other. The EEG signal was also recorded simultaneously in both participants and time-frequency analyzes were performed and compared based on two conditions: a convergence condition, which was confirmed when participants' acoustic patterns (see Methods section in Mukherjee et al., 2018 for more details) exceeded a certain similarity threshold on at least two consecutive words; and a non-convergence condition. Three periods of analysis were targeted: before the production of speech (with reference to the speaker), before listening to the speech (with reference to the listener) and during listening to the speech (with reference to the listener). Thus, the main results of the study showed that phonetic convergence was associated with desynchronization in the alpha/low-beta frequency bands at the frontal level during speech preparation and high-beta suppression before and during speech listening at the centro-parietal and frontal levels, respectively. This type of study remains an exception among research carried out in hyperscanning, but it represents a good alternative for studying convergence phenomena and their neural correlates in interactive contexts with a good degree of control.

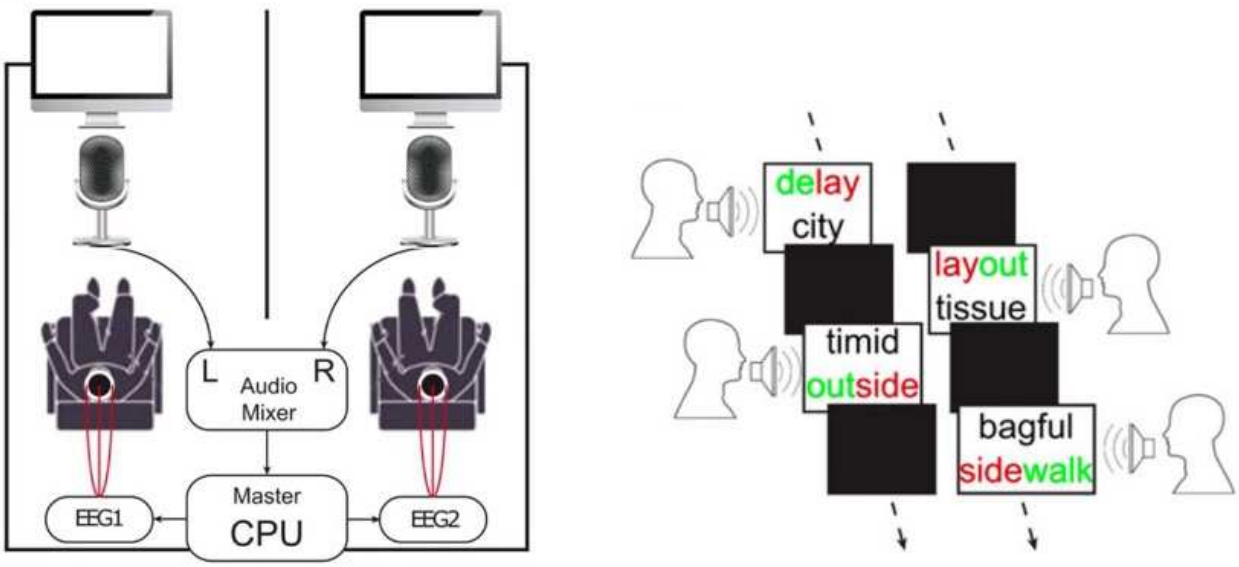


Figure 16 – Set up and Domini task paradigm. *Adapted from Mukherjee et al., 2018.*

c. The case of synchronous speech

Another type of interaction must also be considered when we talk about joint activities, namely synchronous speech, a paradigm which is central in the present thesis work. Initially supported mainly by the studies of Fred Cummins (2002), this paradigm proved to be very interesting for studying an unusual but ecological and realistic situation on the one hand, and on the other hand for studying convergence phenomena. In a synchronous speech task two individuals read a text aloud simultaneously and in synchrony. In real life this situation represents a very specific context for the use of language both in its perception and in its production. Indeed, we can find the production of speech in unison by groups of individuals in worship places (prayers), schools, during birthday songs, in stadiums during sporting events, demonstrations, military parades, political rallies or even in choirs and during musical activities. In all these contexts these are very conventional situations where the prosody used is quite stylized. Synchronous speech is therefore not completely absent in our life, however it is not a priori the norm, nor useful from a communication point of view, where the structure of a conversation is organised by the alternation of the speech turns of the interlocutors.

The first major studies focusing on synchronous speech (Cummins, 2002, 2003, 2004, 2009) highlighted several essential points. It is a task that is remarkably performed without the need of special training and can be performed between a speaker and a recording but also between two speakers. Specifically, across all the studies carried out, the synchronous speech tasks were conducted using the first paragraph of the Rainbow text, displayed on a screen in front of the participants and reorganised by sentence to facilitate reading :

“When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

The rainbow is a division of white light into many beautiful colours.

These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon.

There is, according to legend, a boiling pot of gold at one end.

People look, but no one ever finds it.

When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.”

Through all these studies, the degree of synchrony between speech signals was measured using dynamic time warping (DTW), a method that involves aligning two speech signals continuously and allows a degree of stretching or compression to be quantified to map or distort one signal to another. In this way, it is possible to assess how well two signals are aligned and therefore achieve a quantitative estimate of the asynchrony between the two.

Several experimental conditions were tested, in the first, speakers had to read this paragraph synchronously between themselves (condition performed in the 2002 and 2003 studies). In a second, speakers had to read this paragraph synchronously with a recording which had been made in a natural and unconstrained reading condition. In a third, speakers had to read the paragraph synchronously with a recording that had been made in a synchronous speech condition with another speaker. Participants had been familiarized with the text before the task began, their own speech was routed to the headphones so that they heard the recording in one ear and their own production in the other. Dichotic listening having shown facilitation for synchronization in the various pilots carried out beforehand. One main result has been replicated across numerous Cummins studies, demonstrating that typical asynchronies are around 40ms, with a slight increase to around 60ms in the beginnings of sentences. In addition, participants were able to synchronize better when the task was performed with a person than with a recording (comparison between the first and second conditions). This result highlighted that synchronization is indeed the result of an active process of adaptation with one's partner rather than imitation with a recording. A result which could be reinforced by showing that synchronization with a recording was facilitated when this recording itself bore the characteristics of adaptation since it was recorded in a synchronous situation, although with a different speaker (with reference to the third condition).

In a more recent study led by Bradshaw & Mcgettigan (2021), convergence phenomena occurring during synchronous speech tasks were also reported. In a first session, participants had to read aloud 50 unique sentences, displayed one by one on a screen. This session allowed to measure the reference fundamental frequency (F0) for each sentence. In a second session, participants had to perform the synchronous speech task with audio recordings of another voice in two different conditions, one with high-pitched recordings and a second with low-pitched recordings, to assess whether participants' fundamental frequency would exhibit convergent changes in ascending and descending directions. These hypotheses were confirmed for each of the conditions, highlighting that convergence phenomena similar to those observed in conversational situations are also established in a more restrictive context such as synchronous speech. Bradshaw & Mcgettigan (2021) place these results in the context of speech motor control, as in the case of synchronous speech there is an influence by external vocal inputs on our own speech. An objective was then to confirm that these fundamental frequency convergence phenomena were indeed mainly induced by the acoustic properties of the voice resulting from the recording and not generated by other types of sensory input such as visual information from the speech. To do this they set up another experiment in which the participants had to produce the same types of sentences as in the previous

experiment, in synchronization with silent videos of another speaker, thus preserving the synchronization aspect of the standard synchronized speech condition while removing the acoustic contribution of the accompanist's voice. Still using recordings with modifications of the fundamental frequency, three conditions were compared: a visual synchronized speech condition only, an audio synchronized speech condition (identical to the first experiment) and a combined audiovisual synchronized speech condition, the latter allowing to check whether seeing and hearing the person with whom we are interacting affects the degree of convergence of the fundamental frequency. Firstly, the results replicated the previous ones obtained in the first experiment, synchronization with a high-pitched recording induced increases in the F0 of vocal productions and decreases in the F0 with a low-pitched recording. Secondly, the authors showed that these convergence effects were absent during synchronization with a silent accompanist and that these effects were not better when the participants simultaneously see and hear the accompanist compared to the effects observed in the standard condition (audio only). Thus, the convergence effects that can be observed during a synchronous speech task are specific to the characteristics of the voice with which it is performed, in this case the fundamental frequency.

Replacing synchronous speech in the framework of speech motor control, through this study Bradshaw & Mcgettigan (2021) demonstrated that we adjust certain characteristics of our speech in response to continuous external vocal input. A parallel can be made with a similar known phenomenon in which disturbances of the auditory feedback of our own voice induce compensatory and adaptive changes in the acoustics of the speaker's vocal productions in proportion to the disturbances applied, as in delayed auditory feedack tasks (Lee., 1950; Black., 1951; Fairbanks., 1955; Stuart et al., 2002; Yamamoto & Kawabata, 2014; Karlin et al., 2021). Synchronous speech then represents an interesting situation for models of speech motor control which generally place the processing of auditory feedback from one's own voice as central in the precise adjustment of vocal productions. According to these models (Tourville & Guenther, 2011) falling within the predictive coding framework (see section 2.a.iii : *Feedback control and predictive coding framework*), the brain would make a comparison between a prediction of the expected auditory feedback from a speech utterance with the actual perceived auditory feedback. This would be fundamental on the one hand to maintain the stability of speech and on the other hand to make possible the real-time correction of any potential sensory error (e.g. auditory feedback task). In synchronous speech, the brain simultaneously receives auditory feedback from our own voice as well as the voice of the partner (speaker or recording). The use of this type of task represents an area of investigation which is still to date neglected and under-represented in neuroscience research (see Assaneo et al., 2019), and it would be interesting to assess how this simultaneous vocal feedback could affect on the one hand the processes of language perception and on the other hand the processes of sensorimotor integration during the motor control of speech.

Rationale / transition

Along the path of this introductory chapter, I outlined how the function of language has been progressively studied in neuroscience research. Speech perception and production were initially investigated separately and in isolated context, neglecting the ecological aspect. Then the relationship between perception and production was investigated, gradually enriching the theoretical models of language and emphasizing on the importance to take into account the interactive context to better understand how prediction processes shape speech perception and production. Gradually, improvements in neuroimaging and neural activity recording techniques have enabled more precise measurement of brain responses with greater fidelity and better spatio-temporal resolution. In a parallel way, advancements in signal analysis methods allowed to analyze larger, richer and more complex data sets.

The present PhD work is in line with this desire to investigate language in interaction contexts where individuals must mutually coordinate and adapt their behaviour in order to achieve a goal, as defined by the joint activity framework. The focus here is on verbal coordination, and the objective of this thesis is to bridge verbal coordination and neural dynamics. In this purpose, we are interested in the synchronous speech paradigm, which has been particularly neglected in neurophysiological studies. This paradigm requires the simultaneous and synchronous production of speech between two individuals and thus represents a good interactive framework with a sufficient level of experimental control. It offers several possibilities for the neurophysiological investigation of both speech perception and production and is an interesting case to consider for models of speech motor control.

The initial step of the project was therefore to develop an interactive synchronous speech task, requiring a speaker to continually adapt his or her verbal productions to achieve and maintain synchronization, which represents verbal coordination. In this purpose we developed a novel paradigm, a real-time coupled-oscillators virtual partner (VP) that allows – through coupling strength parameters – to modulate its ability to synchronize speech with a speaker. This leading to a dynamic mutual adaptation of verbal productions.

The second step was to investigate the relation between speech coordination and neural dynamics. In this purpose, we leveraged on the excellent spatiotemporal sensitivity of stereotactic electroencephalography (sEEG) recordings from 16 patients with drug-resistant epilepsy while they performed the interactive synchronous speech task.

In the present work, in one hand we try to address the potential and interest to use a fully dynamic, adaptative and interactive language task to gather deeper understanding of the subtending neural dynamics involved in speech perception, production as well as their interaction. On the other hand, we try to estimate the use of such a task in clinical investigation, as it assesses a more ecological use of language.

Chapter 2 – An interactive synchronous speech task

This chapter presents the first behavioural study carried out during the first year of my PhD thesis. It introduces the task of the project and the experimental paradigm that was developed.
However, this work has not been published in the form of a scientific article due to technical limitations linked to the computer equipment used.

1. How to study verbal coordination in interactive context?

The first stage of my thesis work consisted of setting up a task to study language in an interactional context and to quantify, with a temporally resolved measure, verbal coordination between speakers during a verbal interaction. Verbal interactions are constructed through precise coordination between many different processes affecting speech communication, both within and between speakers (Pickering & Garrod, 2004 ; see *Chapter 1. II. The language in interaction*). Verbal coordination can be difficult to define in a simple way as it involves many heterogeneous contextual factors referring to several domains such as physiological, phonological, pragmatic, semantic, lexical or emotional. Thus, to control the factors that govern the evolution over time of a verbal interaction we must be able to manipulate the behaviour of the interacting partners in real time.

At this aim, the work initially developed by Leonardo Lancia (Lancia et al., 2017) was a fundamental building block of my thesis. It consists of a platform for studying the temporal coordination between speakers involved in verbal interactions. The platform uses a dynamical system model to measure and model the behaviour of a human speaker and, by manipulating model's parameters, to control the behaviour of a virtual partner (VP), producing speech simultaneously with the human speaker. Modelling the behaviour of a human speaker implies predicting at any given moment how it will behave in the near future. To simplify the model, we only focus on interactions wherein the speech content is known in advance (i.e. specific utterances). By identifying at each moment the portion of speech produced by the speaker, we can predict its evolution in the near future and manipulate the VP's behaviour accordingly. Ultimately, this makes it possible to manipulate when, with respect to the real partner, the VP will produce every single element of the utterance.

2. The behavioural study

The task used in my thesis work is an extension of the one designed in a previous study by Lancia et al. (2017). To a certain extent it generalizes to verbal interactions the human dynamical clamp paradigm introduced by Dumas et al. (2014), and limited to finger movements. In this work, the virtual partner is driven by a coordination law as defined in the Haken-Kelso-Bunz model (HKB ; Kelso, 2008) and interacts in a cyclic task with a human partner. Similarly, Lancia and colleagues (2017) modelled speech utterances as sequences of syllabic cycles and coordinated the behaviour of the two partners as coupled oscillators at the syllaboc level. Specifically, the task consisted of the synchronized repetition of a single short utterance (/topkop/) between the human speaker and the virtual partner during 14 seconds (representing one trial). The key point was the virtual partner's ability to adapt its productions in real time to those of the

speaker. The details of the model and the task will be described in the following sections through the new version of the task used for my thesis work.

Thus, the aim of this first behavioural investigation was to : 1) adapt the task used in the study by Lancia et al. (2017) by using more ecological stimuli ; 2) assess the parameters having an impact on verbal coordination.

a. Materials and methods

Participants

30 participants (17 women, mean age 24.7 y, range 19-42 y) took part in the study. All were French native speakers with normal hearing and no neurological disorders. Participants provided written informed consent prior to the experimental session. 5 participants (2 women) were excluded from analysis for poor signal-to-noise ratio in speech recordings.

Data acquisition

The speech signal was recorded using a headset microphone (AKG MICROMIC C520) positioned close to the participants' lips corner. Standard headphone was used for sound presentation. The parameters and sound adjustment were set using an external low-latency sound card (RME Babyface Pro Fs), allowing a tailored configuration for each participant. A calibration was made to find a comfortable volume and an optimal balance for both the sound of the participant's own voice, which was fed back through the headphones, and the sound of the stimuli.

Stimuli

Four stimuli corresponding to four short sentences were pre-recorded by a woman and a man. This allowed to adapt to the natural gender differences in fundamental frequency. All stimuli were normalised in amplitude. Stimuli consisted of four sentences: "papi mamie" (/*papi mami*/, grandpa grandma) / "papi m'a dit" (/*papi ma di*/, grandpa told me) / "mamie lavait ma main" (/*mami lavɛ ma mɛ̃*/, grandma washed my hand) / "mamie manie ma main" (/*mami mani ma mɛ̃*/, grandma handles my hand). The four sentences purposely differed in terms of number of syllables (4-4-6-6). Moreover two of them contained deviations from an otherwise repeating phonological pattern : both in /*papi ma di*/ and in /*mami mani ma mɛ̃*/ , the repeated opening/closing of the lips is substituted by the formation and release of a tongue constriction. These manipulations made the sentences more or less easy to articulate (easy-medium-easy-medium). Sentence durations were 2.07, 2.11, 3.16 and 2.87, respectively.

Experimental design

Participants, comfortably seated on a chair in a soundproof room, performed a synchronous speech task with an artificial agent (Virtual Partner, henceforth VP, see next section) who can modulate and adapt to the participant's speech in real-time.

The experiment required 3 steps (see Figure 1). In a first step participants had to pronounce each sentence twice at natural (spontaneous) speech rate (without the VP). This step was necessary to record the reference fundamental frequency (F0) of each participant specifically for each sentence in order to z-score the stimuli and thus obtain reference stimuli with a fundamental frequency adjusted to each participant. Second, a training phase took place. This required repeating each stimulus over and over together with the VP for ~14 seconds. More precisely, the sentence was first presented on a screen. When the participant pressed the "space" bar, the visual stimulus disappeared and the VP started to "speak". The participant was instructed to repeat the stimuli as synchronously as possible with the VP for the whole trial duration. In the training phase the VP did not adapt to the participant. The training allowed participants to familiarize with the synchronous speech task but also to build a personalised virtual partner model incorporating the articulatory variability of each participant (see section below). The third step was the actual experiment. This was identical to the training but consisted of 48 trials (14s long). Importantly, the VP varied its coupling behaviour to the participant. More precisely, 8 coupling values (noted as k) were tested :

- for 8 sequences the VP had a strong "cooperative" coupling, meaning that it strongly tends to synchronize to the participant speech (k = -0.15).
- for 8 sequences the VP had a moderate "cooperative" coupling (k = -0.09).
- for 8 sequences the VP had a weak "cooperative" coupling (k = -0.03).
- for 8 sequences the VP had a strong "competitive" coupling with a phase shift of pi/2, meaning that it strongly aimed to speak in between the participant syllables (k = +0.15).
- for 8 sequences the VP had a moderate "competitive" coupling (k = +0.09).
- for 8 sequences the VP had a weak "competitive" coupling (k = +0.03).

Moreover, in order to assess the effect of speech rate, we used a slow and fast pace (~3 Hz and ~4 Hz), each stimulus being presented twice. Thus, the final experimental design was organised as follow : 4 stimuli * 6 k values * 2 speech rate, resulting in 48 trials/participant.

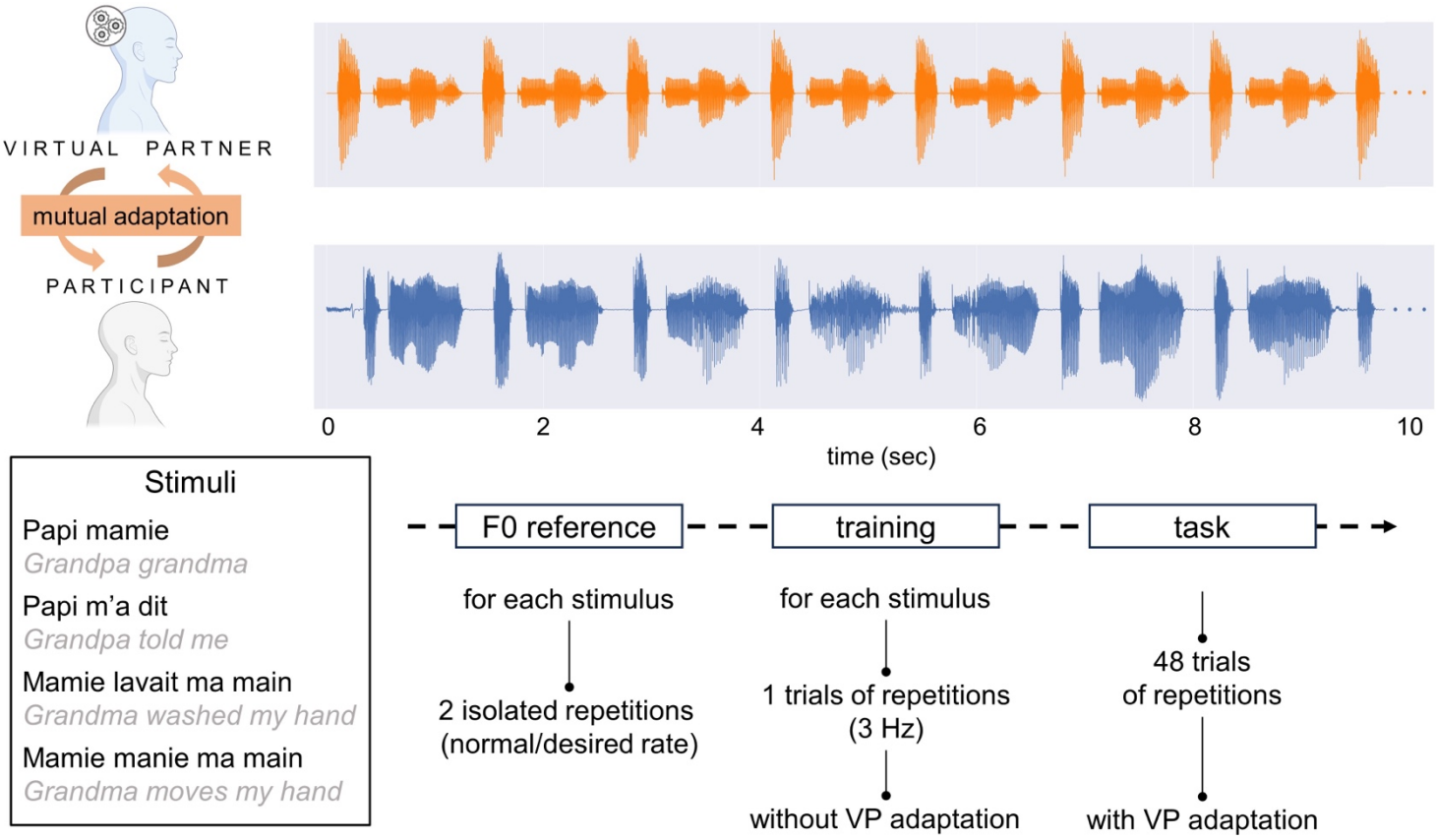


Figure 1 – Task and experimental design. Illustration of one trial of the interactive synchronous speech repetition task (orange : virtual partner speech; blue: participant speech; stimulus « Papi m'a dit » repeated 10 times ; only the 10 first seconds are represented).

Data analysis

Speech signals of the participants and the VP were processed using custom Praat, Matlab and R scripts.

To quantify the degree of coordination of the verbal interaction (verbal coordination index, VCI) we computed the phase locking value for each trial on speech temporal syllabic envelope (the phase was computed using the Hilbert transform), resulting in 48 verbal coordination indexes (VCI) per participant. All statistical analyses were conducted using linear mixed-model with the R package lme4 (Kuznetsova et al., 2017) to take into account the individual variability :

- K effect : *lmer (VCI ~ K + (1|participant))*
- Speech rate effect : *lmer(VCI ~ Speech Rate + (1|participant))*
- Stimuli effect : *lmer(VCI ~ Stimuli + (1|participant))*

Statistical models were compared to a null model : *lmer(VCI ~ 1 + (1|participant))* and post-hoc comparisons were corrected for multiple comparisons using the Tukey test (lsmeans R package ; Lenth, 2016).

b. The virtual partner

The virtual partner (henceforth VP) used for the experiment allows to generate speech (words or short utterances) while adapting it in real-time to the concurrent speech input. The VP environment is built upon the Psychtoolbox-3 program and operates within MATLAB, leveraging custom C subroutines to boost its performance. Its operation revolves around a loop whose iterations are executed at consistent intervals of $\Delta t = 4$ ms. During each iteration, the program analyzes the latest segment (25 ms) of speech produced by the participant and streams a portion of speech to the output device. More precisely, at each iteration of the main loop, the functioning of the VP can be described in four steps:

1. A feature vector is extracted from the last chunk of the input signal.

2. The phase value of the input signal chunk is calculated by mapping the input feature vector onto the corresponding vector of the stimulus signal and retrieving the associated phase value. This step is performed using a dynamic-time-warping algorithm (Dixon, 2005). To enhance precision, the input chunk is mapped onto several model utterances (all time-aligned with the signal used as a stimulus) that are tailored to the characteristics of the participant's speech in the training phase.

3. A chunk of stimulus signal is chosen to be sent to the output device. The selection is guided by applying the Kuramoto (1975) equation to the difference between the phase values representing the current positions of the participant and the VP in their syllabic cycles. This enables real-time lengthening or shortening of speech chunks as needed to adjust towards the preferred phase (0 or π/2). Significantly, in applying this equation, we assign a specific value to the coupling strength parameter « k », linking the behavior of the VP to that of the participant. Values can vary from trial to trial and could be close to 0 (k = +/- 0.03, competitive/cooperative) resulting in a weak coupling behavior (close to a neutral behaviour), negative values (k = -0.09/-0.15) equivalent to a moderate/strong coupling behaviour (tendency to synchronize), or positive values (k = + 0.09/+0.15) corresponding to a moderate/strong coupling behavior with a phase shift of π/2 (tendency to speak in between participant' syllables).

4. The selected chunk of stimulus signal is integrated into the output stream via WSOLA synthesis (Waveform Similarity Overlap-Add).

Note that the adopted modelling schema is formally identical to the one adopted in Assaneo & Poeppel (2020) to model perception action coupling during speech communication.

c. Results

The interactive synchronous speech task enabled us to investigate verbal coordination in an interactive and dynamic context using various parameters applied to the virtual partner (VP) model : the stimuli, the coupling strength (K), and the speech rate.

Verbal coordination index (VCI)

The coordinative behaviour represented by the VCI was affected by the coupling strength parameter K of the model (see Figure 2) with a significant drop in coordination when the VP was parameterised to speak with moderate and strong phase shifts (K=0.09 and K=0.15, respectively) compared to all other K values. This effect was present for each stimulus (see Figure 2 bottom). Surprisingly, we also found a smaller but significant drop when k was strongly set towards a cooperative behaviour (k = -0.15). Table 1 summarises in detail all the results for each comparison between values assigned to the coupling strength parameter.

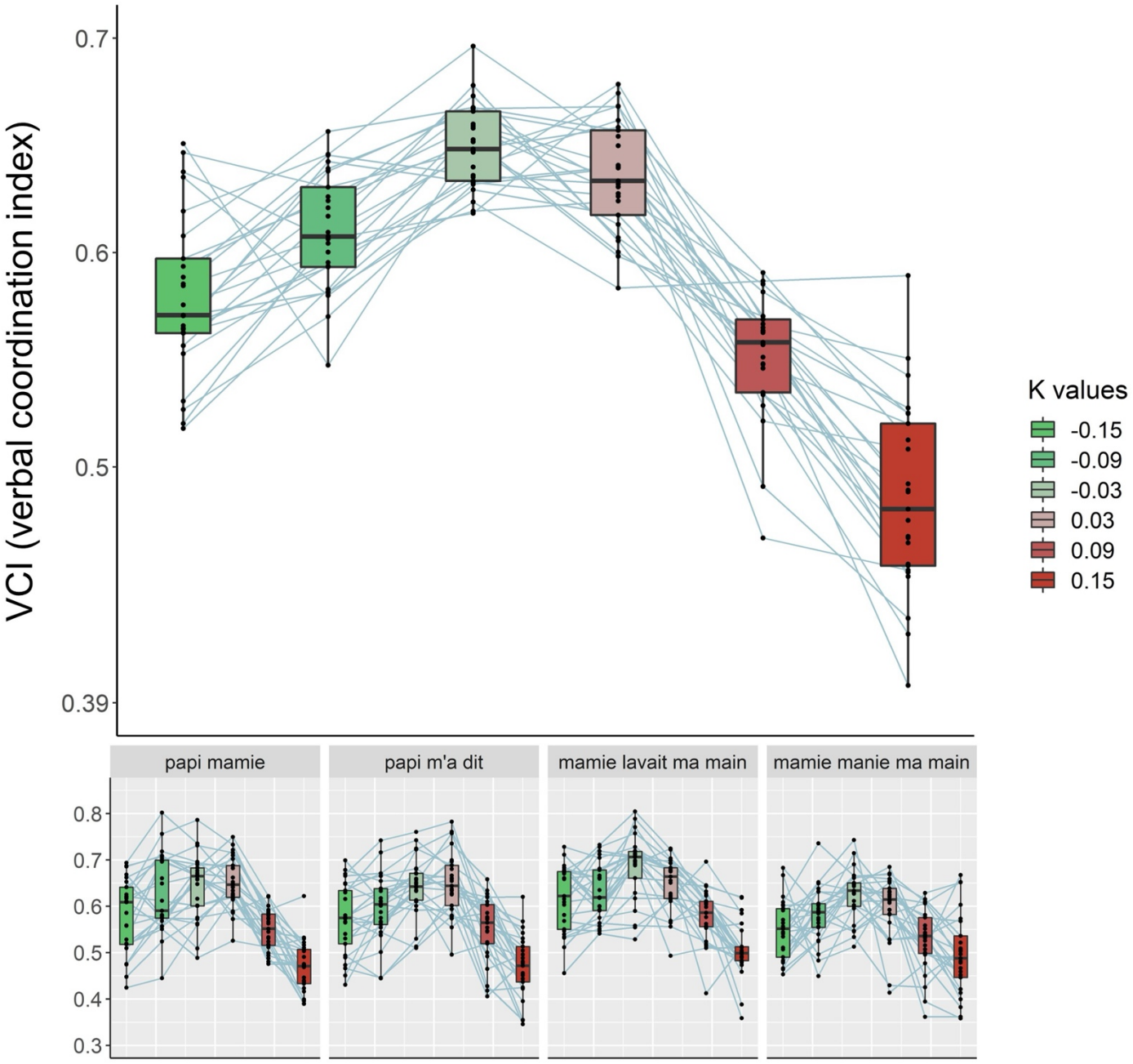


Figure 2 – Verbal coordination index (VCI) according to the coupling strength parameter (K). Top : illustration of the VCI in relation to each K values across all stimuli. The results of the statistical tests are shown in Table 1 (see below). Bottom : representation of the VCI for as a fonction of K for each separate stimulus. The light blue lines in all graphs illustrate the correspondences between the conditions for each participant.

Contrast	t. ratio	p.value
0.09 – 0.15	7.10	<.0001
0.03 – 0.15	16.08	<.0001
(-0.03) – 0.15	17.58	<.0001
(-0.09) – 0.15	13.30	<.0001
(-0.15) – 0.15	10.08	<.0001
0.03 – 0.09	8.98	<.0001
(-0.03) – 0.09	10.48	<.0001
(-0.09) – 0.09	6.20	<.0001
(-0.15) – 0.09	2.98	0.03
(-0.03) – 0.03	1.49	0.66
(-0.09) – 0.03	-2.78	0.06
(-0.15) – 0.03	-6	<.0001
(-0.03) – (-0.09)	4.28	0.0003
(-0.03) – (-0.15)	7.50	<.0001
(-0.09) – (-0.15)	3.21	0.016

Table 1 – Linear mixed-model results for VCI ~ K. The left column reports the comparisons conducted between the different values of the coupling strength parameter K (+/- 0.03, 0.09, 0.15). The middle and right columns show the corresponding statistical results, the t.ratio and the p-value, respectively.

The different stimuli used affected the coordinative behaviour (VCI) (see Figure 3), with the "Mamie lavait ma main" stimulus generating the highest coordination with a significant difference compared to the other three stimuli. In addition, the "Mamie manie ma main" stimulus manifested a significant lowest coordination compared to "Papi, mamie". Table 2 illustrates a detailed summary of the results for each comparison. Finally, modification of speech rate had a significant effect on VCI , with better coordination with the slower pace (3Hz) (t.ratio=11.34, p=<.0001).

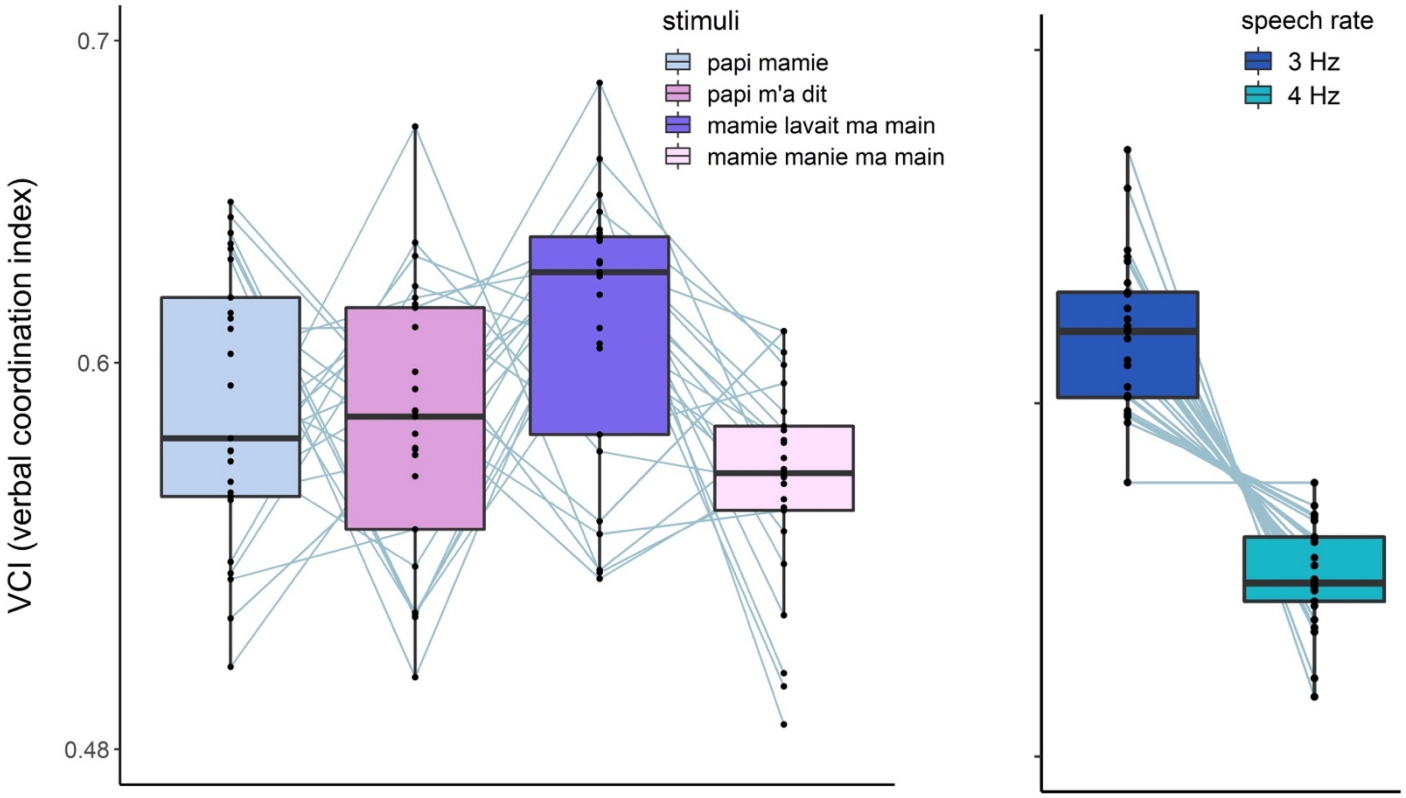


Figure 3 – Verbal coordination index (VCI) according to stimuli and speech rate. The graph on the left illustrates the VCI according to the different stimuli. The graph on the right illustrates the VCI according to the speech rate. The scales in the two graphs are identical and the light blue lines illustrate the correspondences between the conditions for each participant.

Contrast	t. ratio	p.value
Papi, mamie – Papi m'a dit	0.66	0.91
Papi, mamie – Mamie lavait ma main	-2.88	0.02
Papi, mamie – Mamie manie ma main	2.68	0.03
Papi m'a dit – Mamie lavait ma main	-3.54	0.002
Papi m'a dit – Mamie manie ma main	2.01	0.18
Mamie lavait ma main – Mamie manie ma main	5.56	<.0001

Table 2 – Linear mixed-model results for VCI ~ stimuli. The left column reports the comparisons conducted between the different stimuli. The middle and right columns show the corresponding statistical results, the t.ratio and the p-value, respectively.

3. Adaptation of the task in a clinical environment

In this section I will discuss how the synchronous speech task was adapted to be performed in a clinical environment by patients with drug-resistant epilepsy as the main study of my thesis work is based on intracranial data. Therefore, the objective here is not to present a conventional discussion with a detailed comparison of the results with the literature, but to explain the choices that were made regarding the selection of the parameters retained for the experimental task with the patients. The format of the task essentially had to meet two criteria : firstly, it had to be neither too simple nor too laborious to be performed by patients ; secondly, it had to performed in less than ~30-35 minutes.

The main investigation concerned the impact of the coupling strength and its orientation applied to the VP on the verbal coordination index (VCI). By assigning different values to the K parameter, underlying the coupling of the VP's behaviour with that of the participant, two opposite coordination contexts resulted : positive values of K (0.15, 0.09, 0.03) causing the VP to be tendentially in phase shift (π/2) ; negative values (-0.15, -0.09, -0.03) causing the VP to be tendentially in phase. The effect of coupling strength for values of 0.09 and 0.15 resulting in a lower VCI was expected given that the VP was actively driven to be in phase shift, the synchronization that the participants had to achieve was therefore more laborious. On the other hand, the negative values of -0.09 and -0.15 did not result in a higher VCI compared with the lowest values of K (-0.03 and 0.03), whereas it was expected that the VCI would be higher given that the VP was actively tending to be in phase with the participants. An initial explanation of this result is that the application of the coupling factor (with K values resulting in moderate or strong coupling) results in the adaptation of the VP to the participant regardless of the coupling direction (in-phase or phase-shifted). In the specific case of an interactive synchronous speech task, there is mutual adaptation between the VP and the participant, which means that whatever the VP's behaviour, the participant will also adapt his or her speech productions to try and keep as synchronized as possible. It is therefore important to consider that the behaviour, initially defined as "cooperative behaviour", may to some extent disturb the participants and thus represent a difficulty in synchronising. A second explanation of this result is technical. Specifically, the computer hardware on which the task was carried out had an impact on the application of the VP model, particularly the processing of the participants' speech signal and the application of the coupling parameter in a real-time manner. Indeed, real-time computing, although very advantageous by providing an active interactive context, is nevertheless subject to the hardware capabilities on which a model is applied, and in particular with regard to the application of in-phase coupling, which is more sensitive to these performance limits than the application of phase-shift coupling. Indeed, in technical and detailed analyses not described above, we realized that the hardware was skipping from times-to-times 20ms due to the heavy computation demand of real-time speech processing. Despite this, the VCI was higher for values of -0.15 and -0.09 compared to the inverse coupling conditions, confirming that the phase-

shift coupling was indeed applied as expected. This problem was corrected in the patient version of the experiment, but prevented us to publish our data of the control experiment, even though the impact is possibly small.

Importantly, for the aim of this study, the inspection of the general distribution of VCI values across all participants' trials shows a wide variability (participants range : 0.24 - 0.89, mean : 0.58, median : 0.58; see Figure 6 below). This was indeed the desired outcome, enabling a wide gradient of speech synchronization to be obtained between the virtual partner and the participants, that, in the case of patients, would allow correlating coordinative behaviour variability with neural activity variability.

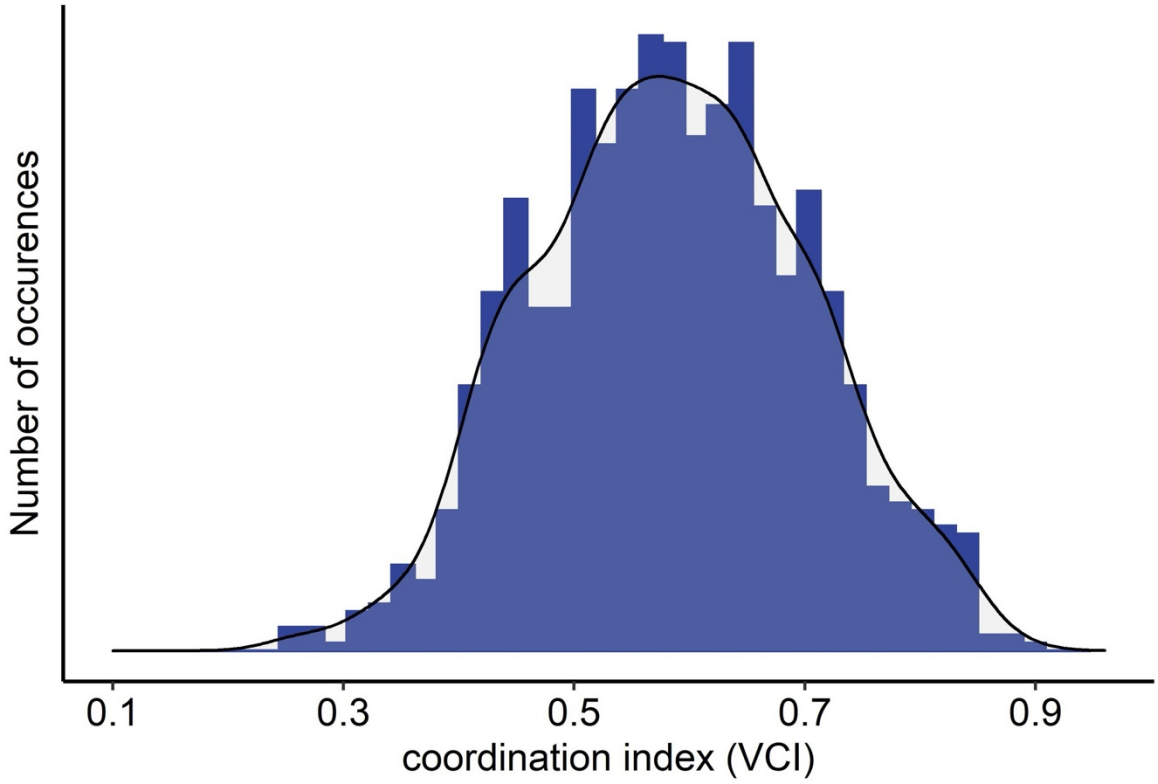


Figure 6 – Distribution of verbal coordination index (VCI) for all participants.

In order to reduce the number of trials and to favour an implicit nature of the virtual partner's behaviour we decided to exclude the extreme coupling values (-0.15 and 0.15). The values leading to moderate coupling in phase and phase shift (-0.09 and 0.09, respectively) make it possible to generate significant differences in the VCI while remaining subtle in terms of the adjustments required to achieve synchronization. These two K values were therefore retained. Finally, the low coupling values (-0.03 and 0.03) showing no difference in VCI were replaced by lower values (-0.01 and 0.01) in order to model control conditions ; values as close to 0 having no impact on the virtual partner behaviour (i.e. no coupling).

The other main parameters were then stimuli and speech rate. With regard to the speech rate, in order to make the task easier for patients, we only kept the speech rate at 3 Hz as it produced the highest VCI. Considering the stimuli choice, although some differences were present between stimuli, as expected due to different articulatory difficulties, we decided to keep the four, again in order to maximize variability of coordinative behaviour across trials. In conclusion, the important parameters retained for the intracranial study were the coupling factor K (+/-0.09, +/-0.01), the four stimuli and the speech rate (3 Hz). A final adjustment was to organise the task into two sessions, each consisting of 24 repetition sequences. This choice was made in order to take into account the fatigue of the patients and to provide a minimum amount of data for the main analyses.

Chapter 3 – Bridging verbal coordination and neural dynamics.

The present chapter corresponds to my **research article** produced during my PhD thesis. This study bridges verbal coordination and neural dynamics, using the interactive synchronous speech task.

It has been submitted to the journal Elife.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bridging verbal coordination and neural dynamics

Isaïh Schwab-Mohamed[1,2], Manuel R. Mercier[1], Agnès Trébuchon[1,3], Benjamin Morillon[1,2], Leonardo Lancia[2,4], Daniele Schön[1,2]

[1]Aix Marseille University, Inserm, INS, Inst Neurosci Syst, Marseille, France.
[2]Aix-Marseille Univ, Institute of Language, Communication and the Brain, France.
[3]APHM, Hôpital de la Timone, Service de Neurophysiologie Clinique, Marseille, France.
[4]Aix Marseille University, CNRS, Laboratoire Parole et Langage (LPL), Aix-en-Provence, France.

***Conflict of interests**:* The authors declare no competing interests.

***Acknowledgments**:* We thank all patients for their willingful participation and all the personnel from the epileptology unit.

***Funding sources**:* ANR-21-CE28-0010 (to D.S), ANR-20-CE28-0007-01 & ERC, SPEEDY, ERC-CoG-101043344 (to B.M), ANR-17-EURE-0029 (NeuroMarseille). This work, carried out within the Institute of Convergence ILCB, was also supported by grants from France 2030 (ANR-16-CONV-0002), the French government under the Programme «Investissements d'Avenir», and the Excellence Initiative of Aix-Marseille University (A*MIDEX, AMX-19-IET-004).

***Author contributions**:* I.S.M, L.L and D.S. designed research; L.L. wrote the code for the experimental setup;I.S.M., L.L, A.T. and M.M. acquired data; I.S.M, L.L and D.S. analyzed data; I.S.M and D.S. wrote the paper; I.S.M, A.T., M.M., B.M., L.L and D.S. edited the paper.

Data availability statement: The conditions of our ethics approval do not permit public archiving of anonymised study data. Readers seeking access to the data should contact Dr. Daniele Schön (daniele.schon@univ-amu.fr). Access will be granted to named individuals in accordance with ethical procedures governing the reuse of clinical data, including completion of a formal data sharing agreement.

Code availability statement: Data analyses were performed using custom scripts in Matlab & Python, and will be available upon publication on Github.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Our use of language, which is profoundly social in nature, essentially takes place in interactive contexts and is shaped by precise coordination dynamics that interlocutors must observe. Thus language interaction is high demanding on fast adjustment of speech production. Here, we developed a real-time coupled-oscillators virtual partner that allows - by changing the coupling strength parameters - to modulate the ability to synchronise speech with a speaker. Then, we recorded the intracranial brain activity of 16 patients with drug-resistant epilepsy while they performed a verbal coordination task with the virtual partner (VP). More precisely, patients had to repeat short sentences synchronously with the VP. This synchronous speech task is efficient to highlight both the dorsal and ventral language pathways. Importantly, combining time-resolved verbal coordination and neural activity shows more spatially differentiated patterns and different types of neural sensitivity along the dorsal pathway. More precisely, high-frequency activity in secondary auditory regions is highly sensitive to verbal coordinative dynamics, while primary regions are not. Finally, the high-frequency activity of the IFG BA44 seems to specifically index the online coordinative adjustments that are continuously required to compensate deviation from synchronisation. These findings illustrate the possibility and value of using a fully dynamic, adaptive and interactive language task to gather deeper understanding of the subtending neural dynamics involved in speech perception, production as well as their interaction.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Language has been most frequently studied by separately assessing perception and production in an isolated context, by contrast with the interactive context that characterizes best its daily usage. The use of language in interactive context, particularly during conversations, calls upon our numerous predictive and adaptive processes (Pickering & Gambi, 2018). Importantly, analyses of conversational data have highlighted the phenomenon of interactive alignment (Pickering & Garrod, 2004), illustrating that during a verbal exchange interlocutors tend to imitate each other and to align their linguistic representations on several levels including phonetic, syntactic, or semantic (Garrod & Pickering, 2009). This is an unconscious and dynamic phenomenon that possibly renders exchanges between speakers more fluid (Marsh et al., 2009). It consists in mutual anticipation (prediction) and coordination of speech production, leading, for instance, to a reduction of turn-taking durations (Levinson, 2016 ; Corps et al., 2018).

Recently, in an effort to assess speech and language in more ecological contexts, researchers in neuroscience have used interactive paradigms to study some of these coordinative phenomena. These consisted of turn-taking behaviors such as alternating naming tasks (Mukherjee et al., 2019), questions and answers investigating motor preparations (Bögels et al., 2015), or manipulating the turn predictability in end-of-turn detection tasks (Magyari et al., 2014 ; for a review see, Bögels et al., 2017). However, the synchronous speech paradigm has been overlooked. This paradigm requires the simultaneous and synchronized production of the same word, sentence, or text between two people. Interestingly, this task is remarkably well performed without any particular training, between a speaker and a recording as well as between two speakers (Cummins, 2002 ; Cummins, 2003 ; Assaneo et al., 2019). As in most joint tasks, individuals must mutually adjust their behavior (here speech production) to optimize coordination (Cummins, 2009). Furthermore, synchronous speech favors the emergence of alignment phenomena, for instance of the fundamental frequency or the syllable onset (Assaneo et al., 2019 ; Bradshaw & McGettigan, 2021 ; Bradshaw et al., 2023). Overall synchronous speech represents a good interactive framework with a sufficient level of experimental control. It offers several possibilities for neurophysiological investigation of both speech perception and production and is an interesting case to consider for models of speech motor control.

Synchronous speech resembles to a certain extent to delayed auditory feedback tasks, which involve real-time perturbations in the speech production signal (such as changes in fundamental frequency and delay). These tasks can induce speech errors as well as modulations in speech and voice features (Stuart et al., 2002 ; Yamamoto & Kawabata, 2014 ; Karlin et al., 2021). Additionally, these tasks provide insights into predictive models of speech motor control, where the brain generates an internal estimate of production and corrects errors when auditory feedback deviates from the estimate (Hickok et al., 2011 ; Houde & Nagarajan, 2011 ; Tourville & Guenther, 2011 ; Ozerk et al., 2022 ; Floegel et al., 2023). Previous studies have revealed increased responses in the superior temporal regions compared to normal feedback conditions (Hirano et al., 1997 ; Hashimoto & Sakai, 2003 ; Takaso et al., 2010 ; Ozerk et al., 2022 ; Floegel et al., 2020). However, synchronous speech paradigms allow investigating the neural bases of coordinative behavior rather than of error correction. Importantly, the precise spectro-temporal dynamics and spatial distribution of the cortical networks underlying speech coordination remain unknown.

To address this issue, we first developed a real-time coupled-oscillators virtual partner that allows - by changing the coupling strength parameters - to modulate the ability to synchronise speech with a speaker. The virtual partner (Lancia et al.,2017) and the synchronous speech task were first tested on a control group to ensure the ability of the virtual agent to coordinate its speech production in real time with the participants. In certain

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interactions, the agent was programmed to actively cooperate with the participants by synchronizing its syllables with theirs. In other interactions, the agent was programmed to deviate from synchronization by producing its syllables between those of the participants. As a result, participants are constantly required to adapt their verbal productions in order to maintain synchronization. Appropriate tuning of the coupling parameters of the virtual agent enabled us to create a variable context of coordination yielding a broad distribution of phase delays between the speaker and agent productions. Then, we leveraged the excellent spatial specificity and temporal resolution of stereotactic depth electrodes recordings and acquired neural activity from 16 patients with drug-resistant epilepsy while they performed the adaptive synchronous speech task with the virtual partner.

Materials and methods

Control – participants

30 participants (17 women, mean age 24.7 y, range 19-42 y) took part in the study. All were French native speakers with normal hearing and no neurological disorders. Participants provided written informed consent prior to the experimental session and the experimental protocol was approved by the Institutional Review board of the French Institute of Health (IRB00003888). 5 participants (2 women) were excluded from analysis for poor signal-to-noise ratio in speech recordings.

Patients – participants

16 patients (7 women, mean age 29.8 y, range 17 - 50 y) with pharmacoresistant epilepsy took part in the study. They were included if their implantation map covered at least partially the Heschl's gyrus and had sufficiently intact diction to support relatively sustained language production. All patients were French native speakers. Neuropsychological assessments carried out before stereotactic EEG (sEEG) recordings indicated that they had intact language functions and met the criteria for normal hearing. In none of them were the auditory areas part of their epileptogenic zone as identified by experienced epileptologists. Recordings took place at the Hôpital de La Timone (Marseille, France). Patients provided written informed consent prior to the experimental session and the experimental protocol was approved by the Institutional Review board of the French Institute of Health (IRB00003888).

Data acquisition

The speech signal was recorded using a microphone (RODE NT1) adjusted on a stand so that it was positioned in front of the participant's mouth. Etymotic insert earphones (Etymotic Research E-A-R-TONE gold) fitted with 10mm foam eartips were used for sound presentation. The parameters and sound adjustment were set using an external low-latency sound card (RME Babyface Pro Fs), allowing a tailored configuration for each participant. A calibration was made to find a comfortable volume and an optimal balance for both the sound of the participant's own voice, which was fed back through the headphones, and the sound of the stimuli.

The sEEG signal was recorded using depth electrodes shafts of 0.8 mm diameter containing 5 to 18 electrode contacts (Dixi Medical or Alcis, Besançon, France). The contacts were 2 mm long and were spaced from each other by 1.5 mm. The placement of the electrode implantations were determined solely on clinical grounds. The cohort consists of 16 unilateral implantations (9 left, 7 bilateral implantations, yielding a total of 236 electrodes and 3185 contacts (see Figure 1). Due to insufficient cover of the right hemisphere we only focus here on the left hemisphere in all presented analyses. All Patients were recorded in a sound-proof Faraday cage using a 256-channels amplifier (Brain Products), sampled at 1kHz and high-pass filtered at 0.16 Hz.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

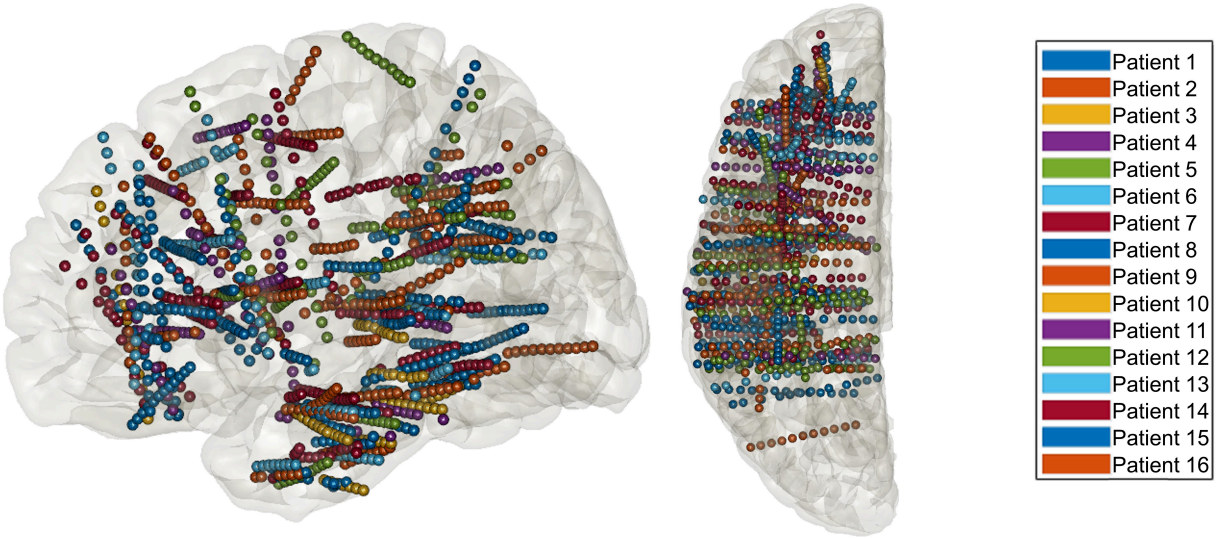


Figure 1 – Anatomical localization of the sEEG electrodes for each patient (N=16).

Stimuli

Four stimuli corresponding to four short sentences were pre-recorded by a woman and a man. This allowed to adapt to the natural gender differences in fundamental frequency. All stimuli were normalised in amplitude.

Stimuli consisted of four sentences: "papi mamie" (/*papi mami*/, grandpa grandma) / "papi m'a dit" (/*papi ma di*/, grandpa told me) / "mamie lavait ma main" (/*mami lavɛ ma mɛ̃*/, grandma washed my hand) / "mamie manie ma main" (/*mami mani ma mɛ̃*/, grandma handles my hand). The four sentences purposely differed in terms of number of syllables (4-4-6-6). Moreover two of them contained deviations from an otherwise repeating phonological pattern : both in /*papi ma di*/ and in /*mami mani ma mɛ̃*/ , the repeated opening/closing of the lips is substituted by the formation and release of a tongue constriction. These manipulations made the sentences more or less easy to articulate (easy-medium-easy-medium). Sentence durations were 2.07, 2.11, 3.16 and 2.87s, respectively with a syllable rate of ~3Hz.

Experimental design

Participants, comfortably seated on a medical chair, performed a real-time interactive synchronous speech task with an artificial agent (Virtual Partner, henceforth VP, see next section) who can modulate and adapt to the participant's speech in realtime.

The experiment required 3 steps (Figure 2A). First, the sentences were presented in written form and the experimenter verified that each participant could pronounce them correctly. Second, a training phase took place. This required repeating each stimulus over and over together with the VP for ~14 seconds. More precisely, the sentence was first presented on a screen. When the participant pressed the "space" bar, the visual stimulus went off, and the VP started to « speak ». The participant was instructed to repeat the stimuli as synchronously as possible with the VP for the whole trial duration. In the training phase the VP did not adapt to the participant.

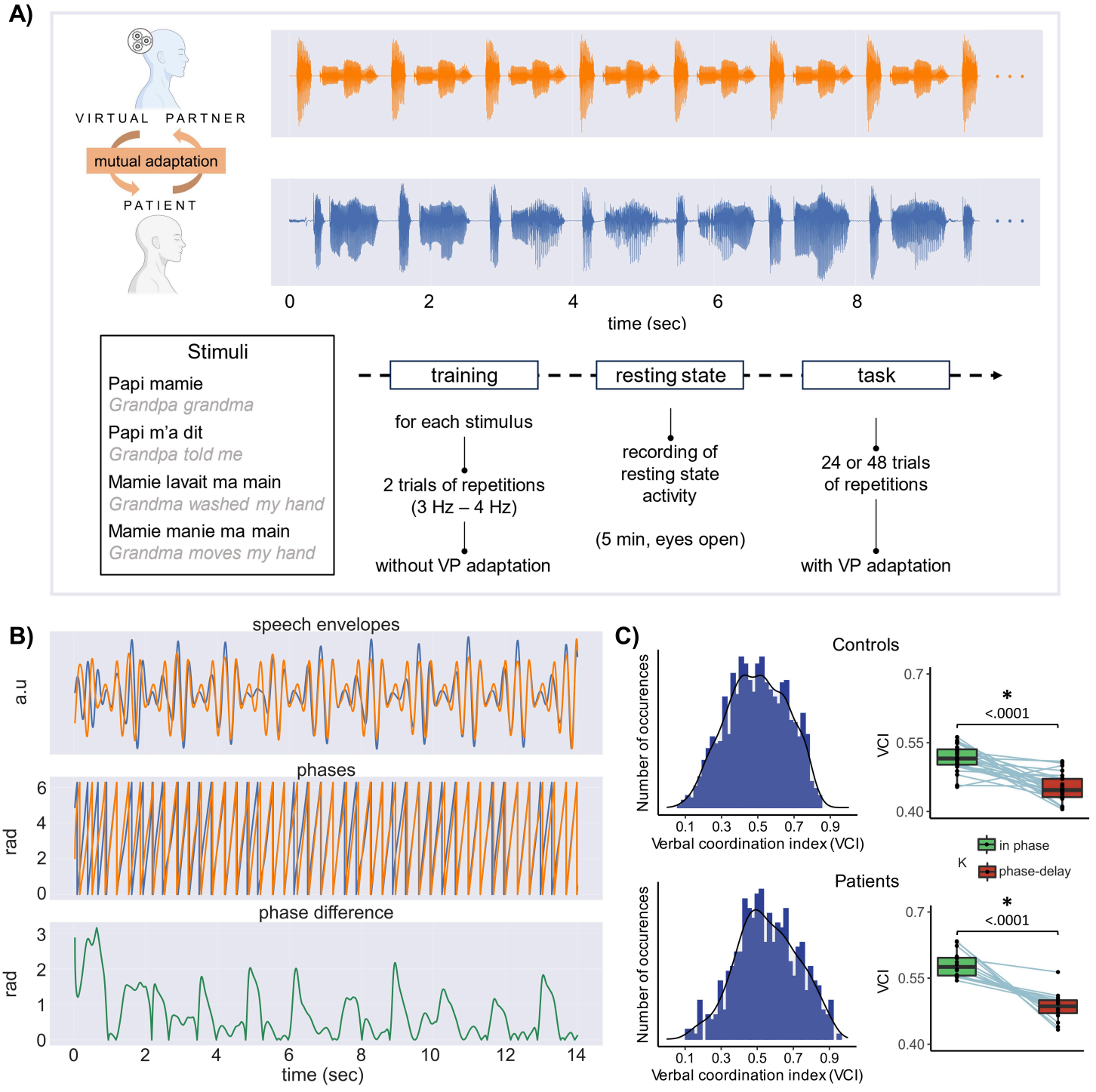
bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The training allowed participants to familiarise with the synchronous speech task but also to build a personalised virtual partner model incorporating the articulatory variability of each participant (see section below).

The training was followed by a resting phase, allowing the recording of resting state activity for a period of 5 min. This time also allowed to build, for each participant, the virtual partner model.

The third step was the actual experiment. This was identical to the training but consisted of 24 trials (14s long, speech rate ~3Hz, yielding ~1000 syllables). Importantly, the VP varied its coupling behaviour to the participant. More precisely, for a third of the sequences the VP had a neutral behaviour (close to zero coupling : k = +/- 0.01). For a third it had a moderate coupling, meaning that the VP synchronised more to the participant speech (k = -0.09). And for the last third of the sequences the VP had a moderate coupling but with a phase shift of pi/2, meaning that it moderately aimed to speak in between the participant syllables (k = + 0.09). Depending on patient fatigue, a second experimental session with 24 extra sequences, was proposed (6/16 patients). Control group participants run a single session of 48 trials.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



177 ***Figure 2 – Paradigm and coordination indexes. (A) Top :*** *illustration of one trial of the interactive synchronous*
178 *speech repetition task (orange: virtual partner speech; blue: participant speech; stimulus papi m'a dit repeated*
179 *10 times ; only the 10 first seconds are represented).* ***Bottom :*** *the four speech utterances used in the task and*
180 *the experimental procedure.* ***(B)*** *Speech signals processing stages. The top panel corresponds to the speech*
181 *envelope, the second to the phase of speech envelope and the third panel to the phase difference between VP*
182 *and participant speech envelopes, illustrating the coordination dynamics along one trial.* ***(C)*** *Left: distributions*
183 *of verbal coordination index (phase locking values between VP and participant speech envelopes, for each trial)*
184 *for all participants (top) and patients . Right: boxplots for control participants (top) and patients showing the*
185 *trial-averaged verbal coordination index as a function of the virtual partner parameters (in-phase coupling vs*
186 *coupling with a 180° shift) .*

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Virtual partner (principles)

The virtual partner (henceforth VP) used for the experiment allows to generate speech (words or short utterances) while adapting it in real-time to the concurrent speech input. The VP environment is built upon the Psychtoolbox-3 program and operates within MATLAB, leveraging custom C subroutines to boost its performance. Its operation revolves around a loop whose iterations are executed at consistent intervals of Δt = 4 ms. During each iteration, the program analyzes the latest segment (25 ms) of speech produced by the participant and streams a portion of speech to the output device. More precisely, at each iteration of the main loop, the functioning of the VP can be described in four steps:

1. A feature vector is extracted from the last chunk of the input signal.

2. The phase value of the input signal chunk is calculated by mapping the input feature vector onto the corresponding vector of the stimulus signal and retrieving the associated phase value. This step is performed using a dynamic-time-warping algorithm (Dixon, 2005). To enhance precision, the input chunk is mapped onto several model utterances (all time-aligned with the signal used as a stimulus) that are tailored to the characteristics of the participant's speech in the training phase.

3. A chunk of stimulus signal is chosen to be sent to the output device. The selection is guided by applying the Kuramoto (1975) equation to the difference between the phase values representing the current positions of the participant and the VP in their syllabic cycles. This enables real-time lengthening or shortening of speech chunks as needed to adjust towards the preferred phase (0 or π/2). Significantly, in applying this equation, we assign a specific value to the coupling strength parameter « k », linking the behavior of the VP to that of the participant. Values can vary from trial to trial and could be close to 0 (k = +/- 0.01) resulting in a neutral behavior, negative values (k = - 0.09) equivalent to a moderate coupling behavior (tendency to synchronize), or positive values (k = + 0.09) corresponding to a moderate coupling behavior with a phase shift of π/2 (tendency to speak in between participant syllables).

4. The selected chunk of stimulus signal is integrated into the output stream via WSOLA synthesis (Waveform Similarity Overlap-Add).

Of note, the coupling strength were chosen to be rather weak and thus do not allow to reach 0 or π/2 phase synchrony, but rather yield the desired large panel of phase delays in the VP-participant coordinative behaviour (see Figure1).

Data analysis

speech signal

Speech signals of the participants and the VP were processed using Praat and Python scripts.

First, raw speech signals were downsampled from 48 KHz to 16 KHz. Then speech envelope was extracted using a pass-band filter between 2.25 and 5 Hz . The phase was computed using the Hilbert transform.

To quantify the degree of coordination of the verbal interaction (verbal coordination index, VCI) we computed the phase locking value using mne_connectivity python function spectral_connectivity_time with the method 'plv' based on the phase coherence model developed by Lachaux et al. (1999) . Phase locking was computed for

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each trial on speech temporal envelope (Fig 1.B), resulting in 24 or 48 verbal coordination indexes (VCI) per patient depending on the number of sessions performed. To assess the effect of the coupling parameter we computed a linear mixed-model contrasting *in-phase coupling trials and trials with a coupling set towards a 180° shift* (lmer(VCI ~ K + (1|participant)). As expected, because of the varying adaptive behaviour of the VP, VCI varies across trials, indexing more or less efficient coordinative behaviour (Figure 1C). Moreover, in order to estimate whether the level of performance was greater than chance level, we computed, for each patient and each trial, a null distribution obtained by randomly shifting the phase between the VP and the patient speech (500 times, see Figure S1).

sEEG signal

General preprocessing related to electrodes localisation

To increase spatial sensitivity and reduce passive volume conduction from neighboring regions (Mercier et al., 2017), the signal was offline re-referenced using bipolar montage. That is, for a pair of adjacent electrode contacts, the referencing led to a virtual channel located at the midpoint locations of the original contacts. To precisely localize the channels, a procedure similar to the one used in the iELVis toolbox was applied (Groppe et al., 2017). First, we manually identified the location of each channel centroid on the post-implant CT scan using the Gardel software (Medina Villalon et al., 2018). Second, we performed volumetric segmentation and cortical reconstruction on the pre-implant MRI with the Freesurfer image analysis suite (documented and freely available for download online http://surfer.nmr.mgh.harvard.edu/). This segmentation of the pre-implant MRI with SPM12 provides us with both the tissue probability maps (i.e. gray, white, and cerebrospinal fluid (CSF) probabilities) and the indexed-binary representations (i.e., either gray, white, CSF, bone, or soft tissues). This information allowed us to reject electrodes not located in the brain. Third, the post-implant CT scan was coregistered to the pre-implant MRI via a rigid affine transformation and the pre-implant MRI was registered to the MNI template (MNI 152 Linear), via a linear and a non-linear transformation from SPM12 methods (Penny et al., 2011), through the FieldTrip toolbox (Oostenveld et al., 2011). Fourth, applying the corresponding transformations, we mapped channel locations to the pre-implant MRI brain that was labeled using the volume-based Human Brainnetome Atlas (Fan et al., 2016).

Signal preprocessing

Continous signal was filtered using 1) a notch filter at 50 Hz and harmonics up to 300 Hz to remove power line artifacts, 2) a bandpass filter between 0.5 Hz and 300 Hz.

To define artifacted channel we used the broadband (raw) signal delimited on the experimental task recording. Channels with a variance greater than 2*IQR (interquartile range, i.e. a non-parametric estimate of the standard deviation) were tagged as artifacted channels (6% of the channels). Channels defined as artifacted were excluded from subsequent analysis.

The continuous signal during the task was then epoched from -0.2 to 14 s relative to the onset of the first stimulus (repetition) of each trial. Generating 24 or 48 epochs depending on the number of sessions each patient had completed. A baseline correction was applied from -0.1 to 0 s. The first 500 ms were cut from the epoched data to avoid the activity burst generated at the beginning of the stimulus.

The continuous signal from the resting state session was also epoched in seventeen 14-second non-overlapping epochs.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Power spectral density

The power spectral density (PSD) computation was conducted for each channels, with the MNE-python function Epochs.compute_psd, trial-by-trial (epochs) in a range of 125 frequencies, logarithmically scaled, ranging from 0.5 to 125 Hz. Six canonical frequency bands were investigated by averaging the power spectra between their respective boundaries: delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), low gamma (30-50 Hz) and high frequency activity or HFa (70-125 Hz).

Global effect

The global effect of the task (versus rest) was computed on each frequency band by susbtracting first the mean resting state activity from the mean experimental activity and then dividing by the mean resting state activity. This approach has the advantage of centering the magnitude before expressing it in percentage (Mercier et al., 2022). For each frequency band and channel, the statistical difference between task activity and the baseline (resting) was estimated with permutation tests (N=1000) using the scipy library.

Coupling behavioural and neurophysiological data

Behavioural speech data and neurophysiological data were jointly analysed using two approaches. In the first approach a two step procedure was used. First, for each frequency band, channel and trial, we computed the mean power. Then, for each frequency band and channel, we computed the correlation across trials using a non-parametric correlation metric (Spearman) between the power and the verbal coordination index (VCI between VP and patient speech). A rho value was thus attributed to each channel and frequency band. The significance was assessed using a permutation approach, similar to the one used for the global effect (see above). In the second approach, we computed the phase-amplitude coupling (PAC) between speech phase and high frequency neural activity. More precisely, we used two regressors for the phase: 1) the phase of the speech envelope of the VP, corresponding to the speech input, 2) the instantaneous phase difference between VP and patient phases, corresponding to the instantaneous coordination of participant and virtual partner (see Figure 2B). As for power, we used the high frequency activity from 70 to 125 Hz. The computation was performed using Tensorpac, an open-source Python toolbox for tensor-based phase-amplitude coupling (PAC) measurement in electrophysiological brain signals (Combrisson et al. 2020).

Clustering analysis

A spatial unsupervised clustering analysis (k-means) was conducted on all significant channels, separately for the global effect (task versus rest) and brain-behaviour correlation analyses. Precisely, we used the silhouette score method on the k-means result (Rousseeuw, 1987 ; Shahapure & Nicholas, 2020). This provides a measure of consistency within clusters of data (or alternatively the goodness of clusters separation). Scores were computed for difference numbers of clusters (from two to ten). The highest silhouette score indicates the optimal number of clusters. Clustering and slihouette scores were computed using the Scikit-learn's Kmeans and silhouette score function (Pedregosa et al., 2011).

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Speech coordinative behaviour

The present synchronous speech task allowed us to create a more or less predictable context of coordination , with the objective of obtaining a wide range of coordination variability. Indeed, the intrinsic nature of the task (speech coordination) on one side and the variable coupling parameter of the VP on the other, require continuous subtle adjustments of the participants speech production. The degree of coordination between speech signals (VP and participant) was assessed at the syllabic level computing the phase locking value between the speech temporal envelopes, a measure of the strength of the interaction (or coupling) of two signals. This metric, the verbal coordination index (VCI), is a proxy of the quality of the performance in coordinating speech production with the VP. Overall, the coordinative behaviour was affected, for both controls and patients, by the coupling parameters of the model with a better coordination when the VP was set to synchronize with participants compared to when it was set to speak with a 180° syllabic shift (controls: t ratio = 4.55, p = <.0001; patients: t ratio = 6.53, p = <.0001, see boxplots in Figure 2C). This produced, as desired, a rather large coupling variability (controls: range : 0.06 - 0.84 ; mean : 0.49 ; median : 0.49 ; patients range : 0.11 - 0.94 ; mean : 0.55 ; median : 0.54, see Fig 2C). This variability was also present at the individual level (see, for patients only, Figure S1). Nonetheless, while variable, the coordinative behaviour was significantly better than chance for every patient (see Figure S1).

Synchronous speech strongly activates the language network from delta to high gamma range

To investigate the sensitivity of synchronous speech in generating spectrally-resolved neural responses, we first analyzed the neural responses in both a spatially and spectrally resolved manner with respect to a resting-state baseline condition. Overall, neural responses are present in all six canonical frequency bands, from the delta range (1-4 Hz) up to high gamma activity (HFa, 70-125 Hz, see Fig 3A) with medium to large modulation (increase or decrease) in activity compared to baseline (Figure 3A). More precisely, while theta, alpha and beta bands show massive desynchronisation, in particular in the STG BA41/42 (primary auditory cortex), STG BA22 (secondary auditory cortex), and IFG BA44 (Broca's area), the low gamma and HFa bands are dominated by power increase in particular in the auditory cortex (STG BA 41/42) and in the inferior frontal gyrus (IFG BA44).

As expected, the whole language network is strongly involved, including both dorsal and ventral pathways (Fig 3A). More precisely, in the temporal lobe the superior, middle and inferior temporal gyri, in the parietal lobe the inferior parietal lobule (IPL) and in the frontal lobe the inferior frontal gyrus (IFG) and the middle frontal gyrus (MFG).

The brain needs behaviour: global vs behaviour-specific neural activity

Comparing the overall activity during the task to activity during rest gives a broad view of the network involved in the task. However, as stated above, the task was conceived to engender a rather wide range of verbal coordination across trials for each participant. This variety of coordinative behaviours allows to explore the link between verbal and simultaneous neural activity, by computing the correlation across trials between the verbal coordination index and the mean power (see methods). This analysis allows us to estimate the extent to which neural activity in each frequency band is modulated as a function of the quality of verbal coordination.

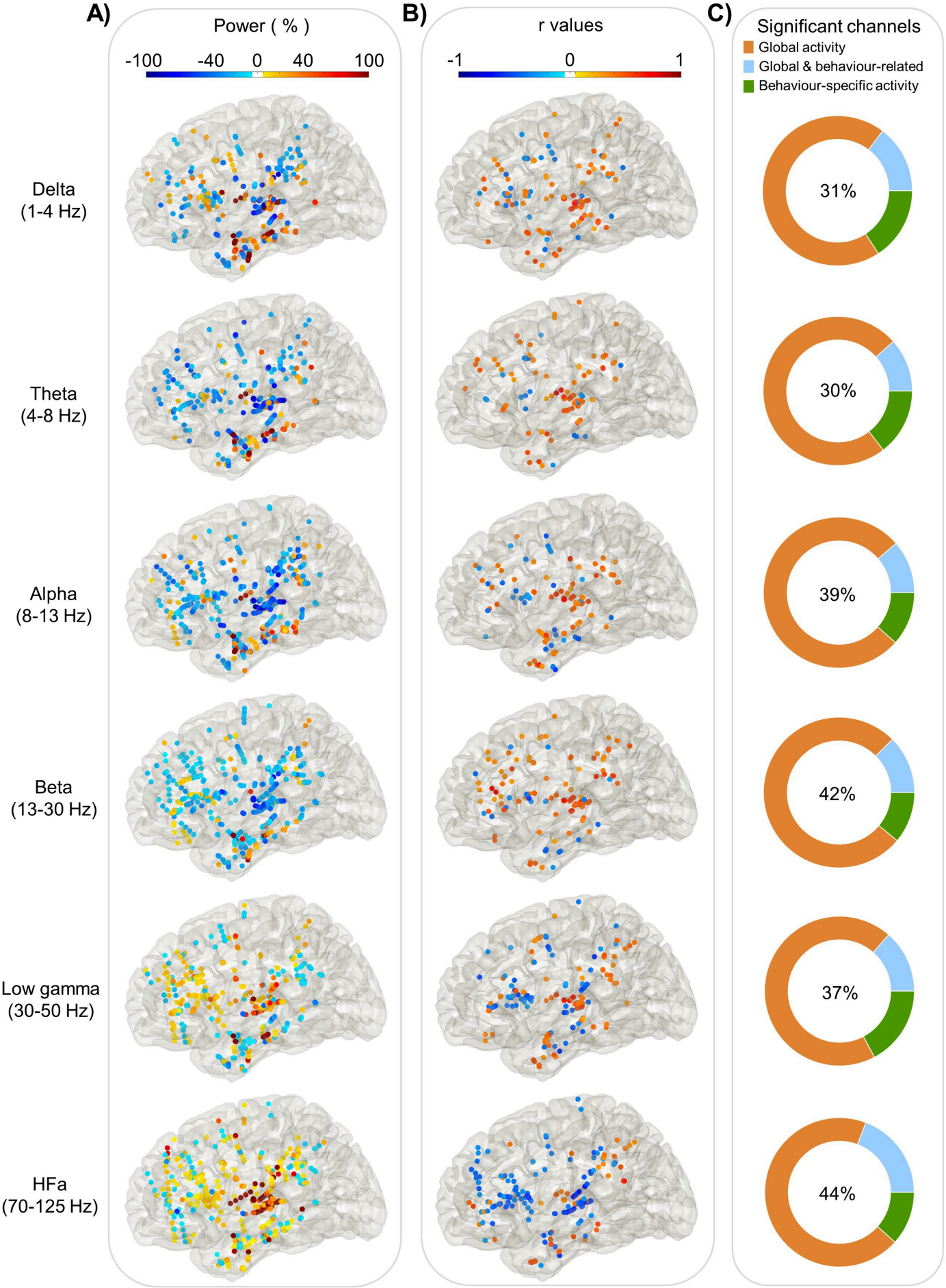
bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3B shows the significant r values (Spearman correlation) for each frequency band. The first observation is a gradual transition in the direction of correlations as we move up frequency bands, from positive correlations at low frequencies to negative ones at high frequencies. This is possibly due to the reversed desynchronization/synchronization process in low and high frequency bands reported above. In other words, while in the low frequency bands stronger desynchronization goes along with weaker verbal coordination, in the HFa stronger activity is associated with weaker verbal coordination.

Importantly, compared to the global activity (task vs rest, Fig 3A), the neural profile of the behaviour-related activity (Fig 3B) is more clustered, in particular in the low gamma and HFa ranges. Indeed, silhouette scores are systematically greater for behaviour-related activity compared to global activity, indicating a better consistency of the clustering independently of the frequency band of interest (see Figure S2). Moreover, the clusters of the behaviour-related activity are most consistent in the HFa range and are located in the IFG BA44, the IPL BA 40 and the STG BA 41/42 and BA22 (see Figure S2).

Comparing global activity and behaviour-related activities shows that, for each frequency band, approximately 2/3 of the channels are only significant in the global activity (Figure 3C, orange part). Of the remaining third, half of the significant channels show a modulation of power compared to baseline that also significantly correlates with the quality of behavioral synchronisation (Figure 3C, blue part). The other half are only visible in the brain-behaviour correlation analysis (Figure 3C, green part, behaviour-specific).

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



361

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3 – Power spectrum analyses. *Each dot represents a significant channel (permutation test).* ***(A)*** *Global activity* ***(****Task versus Rest) for each frequency band. The activity is expressed in % of power change compared to resting.* ***(B)*** *Behaviour-related activity: r values of the Spearman correlation across trials between the iEEG power and the verbal coordination index (VCI).* ***(C)*** *The proportion of significant channels in the task vs rest (orange), in the brain-behaviour correlation (green) and in both (blue). The percentage in the center indicates the overall proportion of significant channels from the three categories (with respect to the total number of channels).*

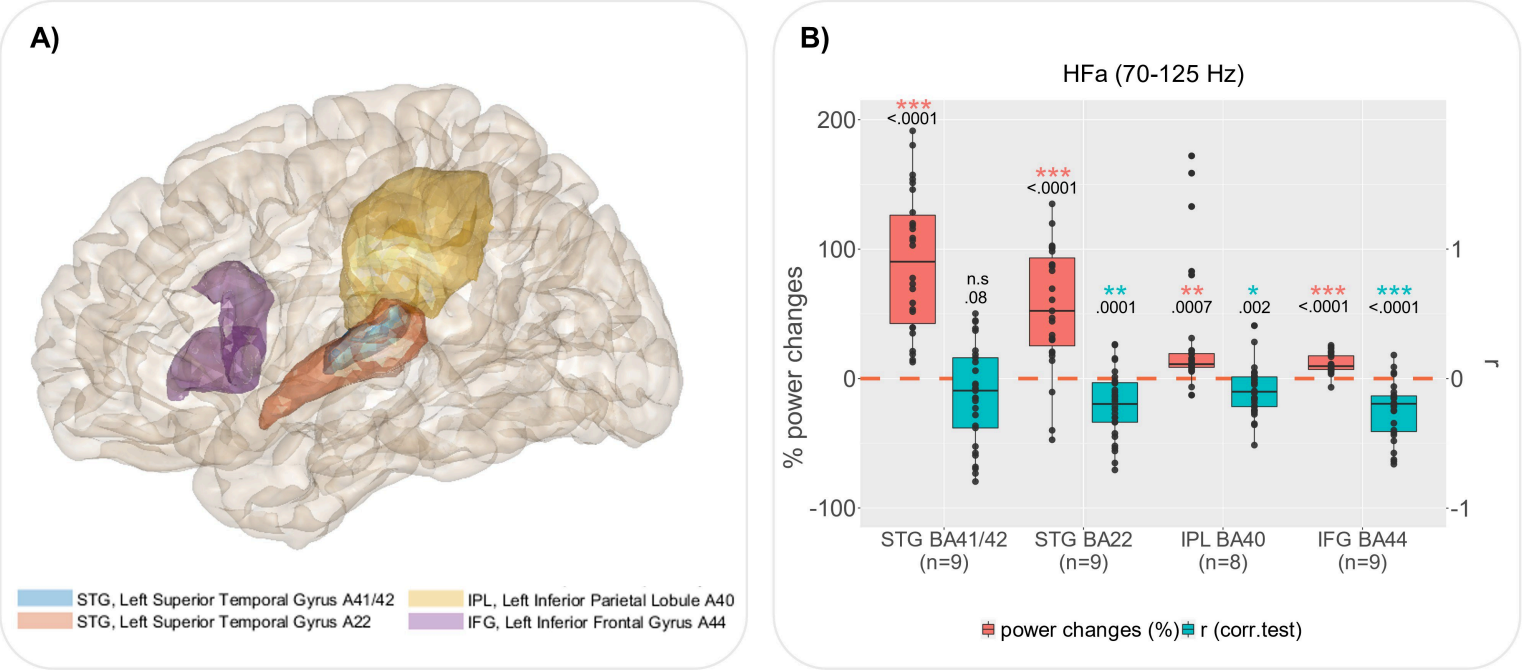


Figure 4 – Group analysis by regions of interest. ***(A)*** *Regions of interest (ROI) defined according to the cluster analysis (see Figure S2), the delimitation of regions is based on the Brainnetome atlas.* ***(B)*** *For each ROI, boxplots illustrate, in red, channels with significant global power changes (HFa, task vs rest) and, in blue, their corresponding r values (correlation between HFa power and verbal coordination index, VCI). Red and blue stars indicate a significant difference from a null distribution. Dots represent independent iEEG channels. The « n » below each region of interest specifies the number of patients. STG : superior temporal gyrus ; IPL : inferior parietal lobule ; IFG : inferior frontal gyrus ; BA : brodmann area.*

Spectral profiles in the language network are nuanced by behaviour.

In order to further explore the brain-behaviour relation in an anatomically language-relevant network, we focused on the frequency band yielding the most consistent clustering (HFa) and on those regions (ROI, in the left hemisphere) recorded in at least 7 patients: STG BA41/42 (primary auditory cortex), STG BA22 (secondary auditory cortex), IPL BA40 (inferior parietal gyrus) and IFG BA44 (inferior frontal gyrus, Broca's region). Within each ROI, using spearman correlation, we quantified the presence and strength of the link between neural activity and the degree of behavioral coordination.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4B shows a dramatic decrease of HFa along the dorsal pathway. However, while STG BA41/42 presents the strongest power increase (compared to baseline), it shows no significant correlation with verbal coordination (t=-1.81, ns). By contrast, the STG BA22 shows a significant negative correlation in the HFa band (t=-4.40, p<0.001), marking a fine distinction between primary and secondary auditory cortex. Finally, the brain-behavior correlation is maximal in the IFG BA44 (t=-5.60, p<0.001).

The IFG is sensitive to speech coordination dynamics.

To model the relation between verbal coordination dynamics and neural activity we conducted a behavior-brain phase amplitude coupling (PAC) analysis at the single trial level, using the power of high frequency neural activity (HFa: 70-125 Hz). For the low-frequency behavioral phase signal, this analysis investigated either the phase of the speech signals of the virtual partner (VP) or the verbal coordination dynamics (phase difference).

When looking at the whole brain analysis (Figure 5A), coupling is strongest as expected in the auditory regions (STG BA41/42 and STG BA22), but it is also present in the left IPL and IFG. Notably, when comparing - within the regions of interest previously described (Figure 5B) - the PAC with the virtual partner speech and the PAC with the phase difference, the coupling relationship changes when moving along the dorsal pathway: a stronger coupling in the auditory regions with the speech input, no difference between speech and coordination dynamics in the IPL and a stronger coupling for the coordinative dynamics compared to speech signal in the IFG (Figure 5B). Similar results were obtained when using the phase of the patient speech rather than the VP speech .

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

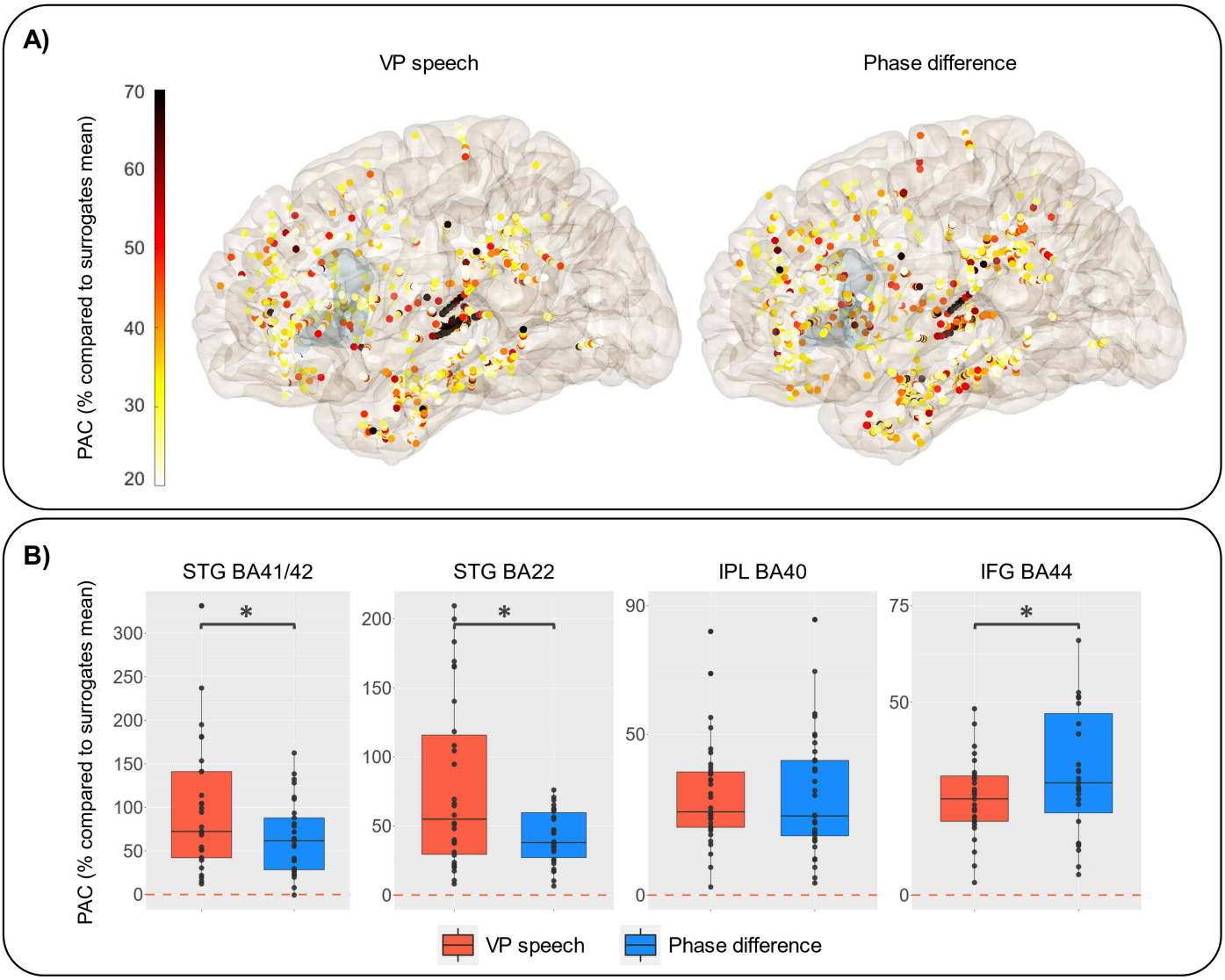


Figure 5 – Phase-amplitude coupling between virtual partner speech signal or coordination dynamics and high frequency activity (HFa). A) *Representation of the increase in PAC expressed in % compared to surrogates when using the virtual partner speech (left) or the coordination dynamics (phase difference between VP and patient, right).The shaded (slight blue) area corresponds to the location of the IFG BA44.* ***B)*** *PAC values for VP (in red) and phase difference (in blue) by regions of interest. Statistical difference calculated using paired wilcoxon test (STG BA41/42 : p=0.01 ; STG BA22 : p=0.004 ; IPL BA40 : p=0.6 ; IFG BA44 : p=0.02). Y-axis range has been adjusted to better illustrate the contrast between VP speech and coordination dynamics.*

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

In this study, we investigated the speech coordinative adjustments using a novel interactive synchronous speech task (Lancia et al., 2017). To assess the relation between speech coordination and neural dynamics, we capitalized on the excellent spatiotemporal sensitivity of human stereotactic recordings (sEEG) from 16 patients with drug-resistant epilepsy while they produced short sentences along with a virtual partner (VP). Critically, the virtual partner is able to coordinate and adapt in real-time its verbal production with those of the participants, thus enabling to create a variable context of coordination yielding a broad distribution of phase delays between participant and VP productions. Several interesting findings can be emphasized. Firstly, the task involving both speech production and perception is efficient to highlight both the dorsal and ventral pathways, from low to high frequency activity (Figure 3). Secondly, spectral profiles of neural responses in the language network are nuanced when combined with behaviour, highlighting the fact that some regions are « simply » involved in the task, in a general manner, while others are truly sensitive to the quality of verbal coordination (Figure 3). Thirdly, high-frequency activity in secondary auditory regions shows a stronger sensitivity to behaviour (coordination success) compared to primary auditory regions (Figure 4). Finally, the high-frequency activity of the IFG BA44 specifically indexes the online coordinative adjustments that are continuously required to compensate the deviation from synchronisation (Figure 5).

Secondary auditory regions are more sensitive to coordinative behaviour.

The increase compared to baseline observed in high frequency activity (HFa) in both the STG BA41/42 (primary auditory cortex) and STG BA22 (secondary auditory cortex) is associated to different cognitive functions. Indeed, when considering whether such activity correlates with the behavioural coordination index – here the phase locking value between the patient and speaker speech – only the STG BA22 shows a significant correlation with a reduced HFa in strongly coordinated trials. This spatial distinction made between primary and secondary auditory regions in terms of their sensitivity to task demands has already been observed in the auditory cortex in particular in the high frequency activity (Nourski, 2017). The use of auditory target detection paradigms – phonemic categorisation (Chang et al., 2011), tone detection (Steinschneider et al., 2014), semantic categorisation tasks (Nourski et al., 2015) – wherein participants are asked to press a button when they hear the target, revealed task-dependent modulations only in the STG (posterolateral superior temporal gyrus or PLST) and not in Heschl's gyrus. In these studies, modulations corresponded to increases in activity in response to targets compared to non-targets, which has been interpreted in terms of selective attention or a bias toward behaviorally relevant stimuli (Petkov et al., 2004 ; Mesgarani & Chang, 2012). Here we extend these findings to a more complex and interactive context. The negative correlation observed in STG BA22 can be interpreted through the prism of internal forward model theory, suggesting that, during actions, a copy of the motor signal is transmitted to sensory regions to be used as a predictive signal to modulate perception and sensory processing. Thereby, the decrease in high-frequency activity as speech coordination increases could be explained by the closer correspondence between the speech signal perceived from the VP and that produced by the patient. This phenomenon has been previously described as speaker-induced suppression (SIS), characterised by a suppression of the neural response corresponding to the speech signal of one's own voice compared to the speech signals of others (Niziolek et al. 2013 ; Kurteff et al., 2023). In the present task, auditory feedback from the on-going speech stream is used to maintain fluent speech production. Moreover, when the two speech signals come close enough in time, the patient possibly perceives them as its own voice. Although this is not in line with previously described absence of SIS during synchronous speech (Jasmin et al., 2016), this echoes more recent findings of Ozerk et al. (2022) who showed, using an auditory feedback task with different fixed delay conditions, a significant response enhancement in auditory cortex that scaled according to the

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

duration of feedback delay. Overall, our result well fit within the framework of prediction errors and error maps (Friston et al., 2020; Tourville & Guenther, 2011), extending the use of synchronous speech to the study of internal forward models of speech in a more dynamic and less strict context.

Inferior frontal gyrus (BA44) as a seat of speech coordination planning in dynamic context.

Our results highlight the involvement of the inferior frontal gyrus (IFG), in particular the BA44 region, in speech coordination. First, trials with a weak verbal coordination (VCI) are accompanied by more prominent high frequency activity (HFa, Fig.4). Secondly, when considering the within-trial time-resolved dynamics, the phase-amplitude coupling (PAC) reveals a tight relation between the low frequency behavioral dynamics (phase) and the modulation of high-frequency neural activity (amplitude, Fig.5B). This relation is strongest when considering the phase adjustments rather than the phase of speech of the VP per se: larger deviation in verbal coordination are accompanied by increase in high-gamma activity. Overall, these findings are in line with the importance of higher-level frontal mechanisms for behavioural flexibility and their role in the hierarchical generative models underlying speech perception and production (Cope et al., 2017). More precisely, they are in line with works redefining the role of Broca's area (BA44 and BA45) in speech production as being more associated to speech planning rather than articulation per se (Flinker et al., 2015 ; Basilakos et al., 2018). Indeed, electrodes covering Broca's area show a greatest activity before the onset of articulation and not during speech production. This has been interpreted in favour of a role at a prearticulatory stage rather than an « on-line » coordination of the speech articulators at least in picture naming (Schuhmann et al., 2009), word repetition (Flinker et al., 2015 ; Ferpozzi et al., 2018) and turn-taking (Castellucci et al., 2022). According to these studies Broca's area may be a « functional gate » at a preacticulatory stage, allowing the phonetic translation before speech articulation (Ferpozzi et al., 2018). Our use of a synchronous speech task allows to refine this view, by showing that these prearticulatory commands are of continuous rather than discrete nature.

In other terms, the discrete (on-off) and ignition-like behaviour of neuronal populations in Broca's area gating prearticulatory commands before speech may be due to the discrete nature of the tasks used to assess speech production. Notably, picture naming, word repetition, word reading and even turn-onsets imply that speech production is preceded by a silent period during which the speaker listens to speech or watches pictures. By contrast, the synchronous speech task requires continuous temporal adjustments of verbal productions in order to reach synchronisation with the virtual partner. Relatedly, the involvement of IFG in accurate speech timing has been previously shown via thermal manipulation (Long et al., 2016).

Of note, temporal adjustements (prediction error corrections) are also needed for fluent speech in general, beyond synchronous speech, and give rise to the rhythmic nature of speech. Temporal adjustments possibly take advantage of the auditory input and are referred to as audio-motor interactions that can be modeled as a coupled oscillator (Poeppel & Assaneo , 2020). Interestingly, during speech perception the coupling of theta and gamma bands in the auditory cortex reflects tracking of slow speech fluctuations to spiking gamma (Morillon et al., 2010 ; Morillon et al., 2012 ; Hyafil et al., 2015; Lizarazu et al. 2019 ; Oganian & Chang, 2019 ; Leonard et al., 2024), similarly to what we describe in the auditory cortex. By contrast, in the inferior frontal gyrus, the coupling in the high frequency activity is strongest with the input-output phase difference (input of the VP - output of the speaker), a metric that reflects the amount of error in the internal computation to reach optimal coordination, which indicates that this region optimises the predictive and coordinative behaviour required by the task. This well fits with the anatomical connectivity that has been described between Broca's and Wernicke's territory via the long segment of the arcuate fasciculus, possibly setting the base for a mapping from speech representations in Wernicke's area to the predictive proprioceptive adjustments processed in Broca's area (Catani & Ffytche, 2005 ; Oestreich et al., 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, while the case of synchronous speech may seem quite far away from real-life conversational contexts, the models describing language interaction consider that listeners covertly imitate the speaker's speech and timely construct a representation of the underlying communicative intention which allows early fluent turn-taking (Pickering & Gambi, 2018 ; Levinson, 2016). Moreover, synchronous speech has recently gained interest in the neuroscience field due to important results showing a relation between anatomo-functional features and synchronization abilities. More precisely, Assaneo and collaborators (2019) used a spontaneous speech synchronization test wherein participants produce the syllable /tah/ while listening to a random syllable sequence of predefined pace. The authors identified two groups of participants (high and low synchronisers) characterised by their ability to naturally synchronise their productions more or less easily with the auditory stimulus. Importantly, the ability to synchronise correlates to the degree of lateralization of the arcuate fasciculus, that connects the inferior frontal gyrus and the auditory temporal regions, high synchronizers showing greater lateralization to the left than the low synchronizers. The more this structural connectivity of the arcuate fasiculus is lateralised in the left hemisphere, the more the activity of the IFG is synchronised with the envelope of the audio stimulus of the SSS-test, during a passive listening task. A major limitation of EEG and MEG studies is that they are very sensitive to speech production artefacts, which is not the case of iEEG. Thus, the full dynamics of speech interaction are difficult to assess with surface recordings. Our findings extend these results in several manners. First, speech production is not limited to a single syllable, but to complex utterances. Secondly, the auditory stimulus is not preset but adapts and changes behaviour in real-time as a function of the dynamic of the « dyad », here patient-virtual partner. Thirdly, and most importantly, iEEG recording allowing speech artefact-free data, we could extend the relation between coordination abilities and anatomical circuitry of the IFG to the neural dynamics of this same region, showing that it plays an important role in the temporal adjustement of speech that are necessary to synchronize to external speech.

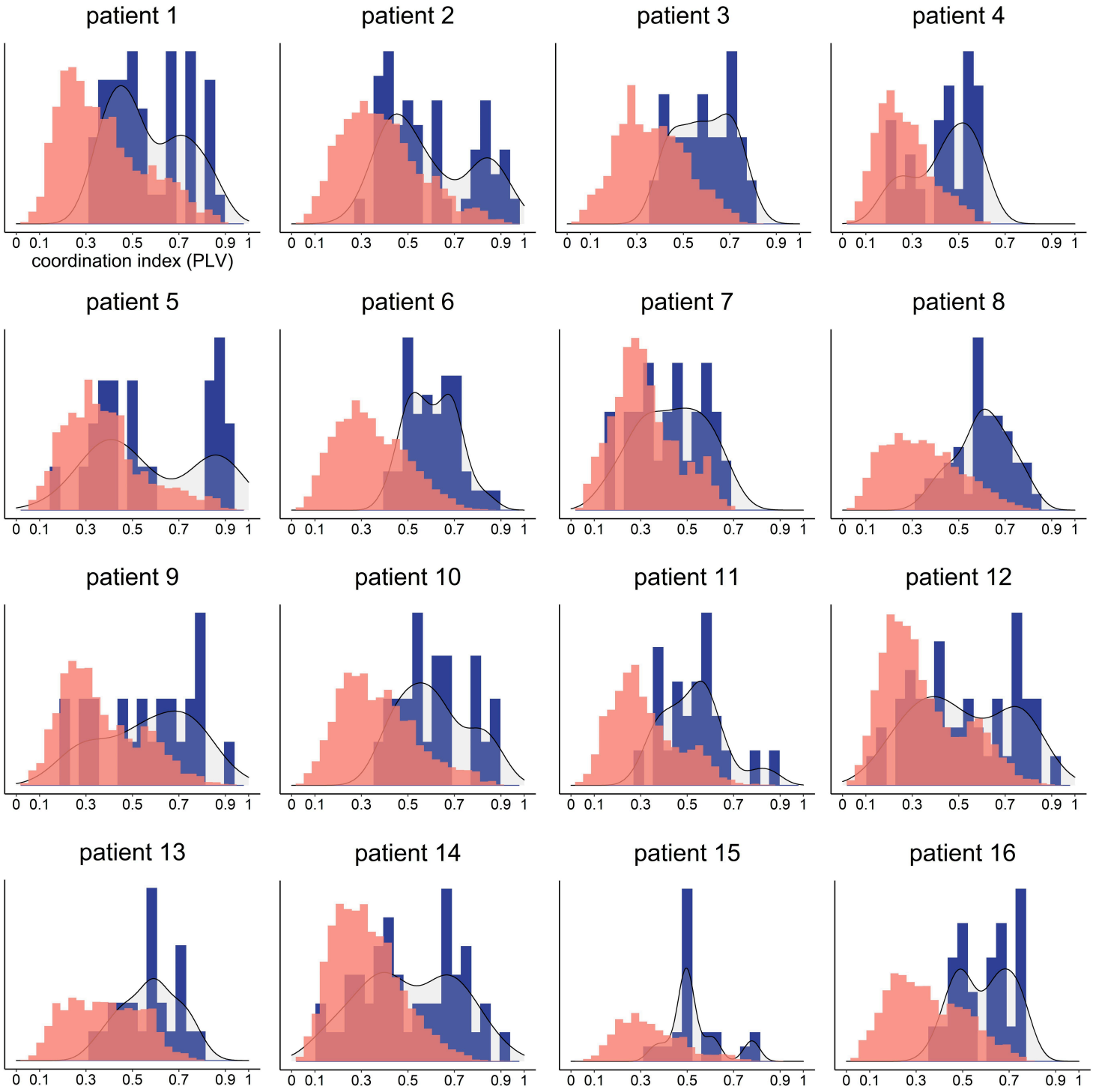
To conclude, the present study illustrates the possibility and interest of using a fully dynamic, adaptive and interactive language task to gather deeper understanding of the subtending neural dynamics involved in speech perception, production as well as their interaction.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

549 **Supplementary Material**

550

551



552

553

554 ***Figure S1 – Distribution of verbal coordination index across patients****. For each of the sixteen patients, this figure*
555 *depicts the histogram of the coordination index for all trials (in blue) as well as the null distribution (random*
556 *phase shift) computed using 500 permutations per trial (in red).*

557

558

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

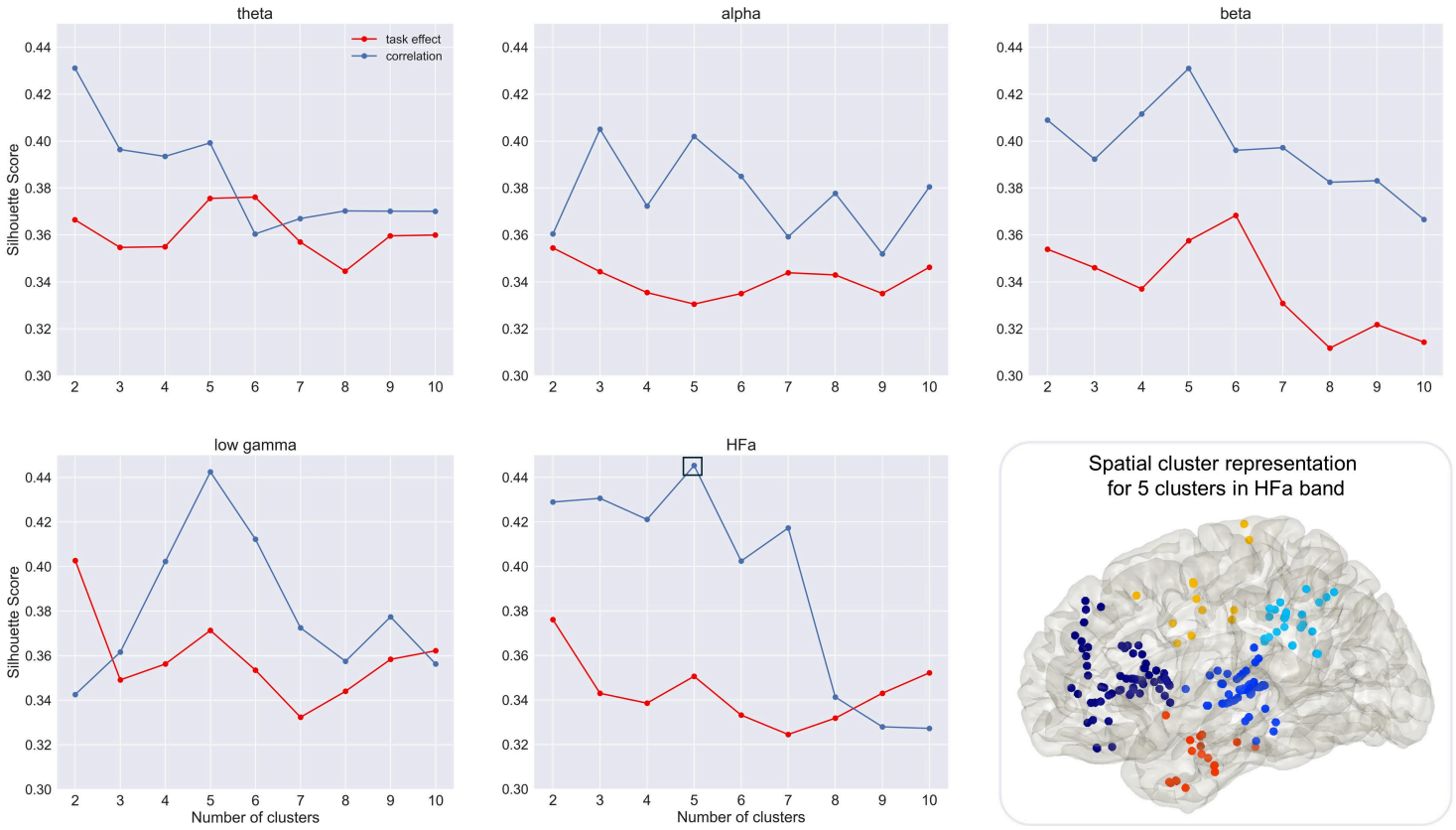


Figure S2 – Cluster analysis (silhouette score). *Illustration of the mean silhouette scores according to the number of clusters for global activity (in red) and behaviour-related activity (correlation between power changes and coordination index, in blue).* The highest silhouette score was obtained for five clusters in the HFa range for behaviour-related activity (framed in full black square). *Bottom right : spatial cluster representation for the highest mean silhouette score value in HFa range.*

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., & Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. Nature neuroscience, 22(4), 627-632.

Basilakos, A., Smith, K. G., Fillmore, P., Fridriksson, J., & Fedorenko, E. (2018). Functional characterization of the human speech articulation network. Cerebral Cortex, 28(5), 1816-1830.

Bögels, S., & Levinson, S. C. (2017). The brain behind the response: Insights into turn-taking in conversation from neuroimaging. Research on Language and Social Interaction, 50(1), 71-89.

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. Scientific reports, 5(1), 12881.

Bradshaw, A. R., & McGettigan, C. (2021). Convergence in voice fundamental frequency during synchronous speech. PloS one, 16(10), e0258747.

Bradshaw, A. R., Lametti, D. R., Shiller, D. M., Jasmin, K., Huang, R., & McGettigan, C. (2023). Speech motor adaptation during synchronous and metronome-timed speech. Journal of Experimental Psychology: General. to articulate the transition.

Castellucci, G. A., Kovach, C. K., Howard III, M. A., Greenlee, J. D., & Long, M. A. (2022). A speech planning network for interactive language use. *Nature*, *602*(7895), 117-122.

Catani, M., & Ffytche, D. H. (2005). The rises and falls of disconnection syndromes. *Brain*, *128*(10), 2224-2239.

Chang, E. F., Edwards, E., Nagarajan, S. S., Fogelson, N., Dalal, S. S., Canolty, R. T., ... & Knight, R. T. (2011). Cortical spatio-temporal dynamics underlying phonological target detection in humans. Journal of cognitive neuroscience, 23(6), 1437-1446.

Combrisson, E., Nest, T., Brovelli, A., Ince, R. A., Soto, J. L., Guillot, A., & Jerbi, K. (2020). Tensorpac: An open-source Python toolbox for tensor-based phase-amplitude coupling measurement in electrophysiological brain signals. PLoS computational biology, 16(10), e1008302.

Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P. S., Wiggins, J., ... & Rowe, J. B. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature communications*, *8*(1), 2154.

Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. Discourse processes, 55(2), 230-240.

Cummins, F. (2002). On synchronous speech. Acoustics Research Letters Online, 3(1), 7-11.

Cummins, F. (2003). Practice and performance in speech produced synchronously. Journal of Phonetics, 31(2), 139-148.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. Journal of Phonetics, 37(1), 16-28.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, *53*(1), 1-15.

Dixon, S. (2005, September). Live tracking of musical performances using on-line time warping. In *Proceedings of the 8th International Conference on Digital Audio Effects* (Vol. 92, p. 97).

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., ... & Jiang, T. (2016). The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, *26*(8), 3508-3526.

Ferpozzi, V., Fornia, L., Montagna, M., Siodambro, C., Castellano, A., Borroni, P., ... & Cerri, G. (2018). Broca's area as a pre-articulatory phonetic encoder: gating the motor program. *Frontiers in human neuroscience*, *12*, 64.

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., & Crone, N. E. (2015). Redefining the role of Broca's area in speech. Proceedings of the National Academy of Sciences, 112(9), 2871-2875.

Floegel, M., Fuchs, S., & Kell, C. A. (2020). Differential contributions of the two cerebral hemispheres to temporal and spectral speech feedback control. *Nature Communications*, *11*(1), 2839.

Floegel, M., Kasper, J., Perrier, P., & Kell, C. A. (2023). How the conception of control influences our understanding of actions. *Nature Reviews Neuroscience*, *24*(5), 313-329.

Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., & Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neuroscience & Biobehavioral Reviews*, *118*, 42-64.

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. Topics in Cognitive Science, 1(2), 292-304.

Groppe, D. M., Bickel, S., Dykstra, A. R., Wang, X., Mégevand, P., Mercier, M. R., ... & Honey, C. J. (2017). iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *Journal of neuroscience methods*, *281*, 40-48.

Hashimoto, Y., & Sakai, K. L. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. Human brain mapping, 20(1), 22-28.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. Neuron, 69(3), 407-422.

Hirano, S., Kojima, H., Naito, Y., Honjo, I., Kamoto, Y., Okazawa, H., ... & Konishi, J. (1997). Cortical processing mechanism for vocalization with auditory verbal feedback. Neuroreport, 8(9), 2379-2382.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. Frontiers in human neuroscience, 5, 82.

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A. L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *elife*, *4*, e06213.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Jasmin, K. M., McGettigan, C., Agnew, Z. K., Lavan, N., Josephs, O., Cummins, F., & Scott, S. K. (2016). Cohesion and joint speech: right hemisphere contributions to synchronized vocal production. *Journal of Neuroscience*, *36*(17), 4669-4680.

Karlin, R., Naber, C., & Parrell, B. (2021). Auditory feedback is used for adaptation and compensation in speech timing. Journal of Speech, Language, and Hearing Research, 64(9), 3361-3381.

Kuramoto, Y. (1975). Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan* (pp. 420-422). Springer Berlin Heidelberg.

Kurteff, G. L., Lester-Smith, R. A., Martinez, A., Currens, N., Holder, J., Villarreal, C., ... & Hamilton, L. S. (2023). Speaker-induced suppression in EEG during a naturalistic reading and listening task. Journal of Cognitive Neuroscience, 35(10), 1538-1556.

Lachaux, J. P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human brain mapping*, *8*(4), 194-208.

Lancia, L., Chaminade, T., Nguyen, N., & Prevot, L. (2017, August). Studying the link between inter-speaker coordination and speech imitation through human-machine interactions. In Interspeech 2017 (p 859-863).

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., ... & Chang, E. F. (2024). Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, *626*(7999), 593-602.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. Trends in cognitive sciences, 20(1), 6-14.

Lizarazu, M., Lallier, M., & Molinaro, N. (2019). Phase− amplitude coupling between theta and gamma oscillations adapts to speech rate. *Annals of the new York Academy of Sciences*, *1453*(1), 140-152.

Long, M. A., Katlowitz, K. A., Svirsky, M. A., Clary, R. C., Byun, T. M., Majaj, N., ... & Greenlee, J. D. (2016). Functional segregation of cortical regions underlying speech timing and articulation. Neuron, 89(6), 1187-1193.

Magyari, L., Bastiaansen, M. C., De Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. Journal of Cognitive Neuroscience, 26(11), 2530-2539.

Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in cognitive science*, *1*(2), 320-339.

Mercier, M. R., Bickel, S., Megevand, P., Groppe, D. M., Schroeder, C. E., Mehta, A. D., & Lado, F. A. (2017). Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: a glimpse on white matter signal. *Neuroimage*, *147*, 219-232.

Mercier, M. R., Dubarry, A. S., Tadel, F., Avanzini, P., Axmacher, N., Cellier, D., ... & Oostenveld, R. (2022). Advances in human intracranial electroencephalography research, guidelines and good practices. *Neuroimage*, *260*, 119438.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233-236.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Morillon, B., Lehongre, K., Frackowiak, R. S., Ducorps, A., Kleinschmidt, A., Poeppel, D., & Giraud, A. L. (2010). Neurophysiological origin of human brain asymmetry for speech and language. *Proceedings of the National Academy of Sciences*, *107*(43), 18688-18693.

Morillon, B., Liégeois-Chauvel, C., Bénar, C. G., & Giraud, A. L. (2012). Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. *Frontiers in psychology*, *3*, 24108.

Mukherjee, S., Badino, L., Hilt, P. M., Tomassini, A., Inuggi, A., Fadiga, L., ... & d'Ausilio, A. (2019). The neural oscillatory markers of phonetic convergence during verbal interaction. Human brain mapping, 40(1), 187-201.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. Journal of Neuroscience, 33(41), 16110-16116.

Nourski, K. V. (2017). Auditory processing in the human cortex: An intracranial electrophysiology perspective. Laryngoscope investigative otolaryngology, 2(4), 147-156.

Nourski, K. V., Steinschneider, M., Oya, H., Kawasaki, H., & Howard III, M. A. (2015). Modulation of response patterns in human auditory cortex during a target detection task: an intracranial electrophysiology study. International Journal of Psychophysiology, 95(2), 191-201.

Oestreich, L. K., Whitford, T. J., & Garrido, M. I. (2018). Prediction of speech sounds is facilitated by a functional fronto-temporal network. Frontiers in Neural Circuits, 12, 43.

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science advances*, *5*(11), eaay6279.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*, 1-9.

Ozker, M., Doyle, W., Devinsky, O., & Flinker, A. (2022). A cortical network processes auditory error signals during human speech production to maintain fluency. PLoS Biology, 20(2), e3001493.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (Eds.). (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

Petkov, C. I., Kang, X., Alho, K., Bertrand, O., Yund, E. W., & Woods, D. L. (2004). Attentional modulation of human auditory cortex. Nature neuroscience, 7(6), 658-663.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. Psychological bulletin, 144(10), 1002.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and brain sciences, 27(2), 169-190.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature reviews neuroscience*, *21*(6), 322-334.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.23.590817; this version posted April 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Schuhmann, T., Schiller, N. O., Goebel, R., & Sack, A. T. (2009). The temporal characteristics of functional activation in Broca's area during overt picture naming. *cortex*, *45*(9), 1111-1116.

Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)* (pp. 747-748). IEEE.

Steinschneider, M., Nourski, K. V., Rhone, A. E., Kawasaki, H., Oya, H., & Howard III, M. A. (2014). Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. Frontiers in neuroscience, 8, 240.

Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. The Journal of the Acoustical Society of America, 111(5), 2237-2241.

Takaso, H., Eisner, F., Wise, R. J., & Scott, S. K. (2010). The effect of delayed auditory feedback on activity in the temporal lobe while speaking: a positron emission tomography study.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. Language and cognitive processes, 26(7), 952-981.

Villalon, S. M., Paz, R., Roehri, N., Lagarde, S., Pizzo, F., Colombet, B., ... & Bénar, C. G. (2018). EpiTools, A software suite for presurgical brain mapping in epilepsy: Intracerebral EEG. *Journal of neuroscience methods*, *303*, 7-15.

Yamamoto, K., & Kawabata, H. (2014). Adaptation to delayed auditory feedback induces the temporal recalibration effect in both speech perception and production. Experimental brain research, 232, 3707-3718.

Chapter 4 – Clinical perspectives

This chapter will focuses on the clinical implications of my thesis work. I think it's important that, whenever possible, fundamental research projects should be considered in light of societal needs. In the case of neuroscience this is often related to a usage in the clinical field. In my thesis work, I had the responsibility and the opportunity to accompany each of the patients during my experimental task, which enabled me to communicate directly with them, the clinical staff, and more particularly with Agnès Trébuchon, neurologist in the epileptology and cerebral rhythmology unit at the Hôpital de la Timone (Marseille, France). The chapter will therefore be organised by first introducing the clinical setting and the patients with drug-resistant epilepsy, then the various clinical tasks used in the department will be outlined and finally a perspective aimed at linking the results of my thesis work with those obtained during the clinical investigations will be proposed. Although this section is somewhat uncomplete, I wished to share it here to be able to present it more thoroughly at my defence.

1. Clinical investigations of drug-resistant epilepsy

Patients suffering from drug-resistant epilepsy may be candidates for neurosurgical procedures aimed at resecting the epileptogenic zone. In order for surgery to be 'successful', this requires to precisely identify not only the (cortical) regions causing the seizures, but also the functional (cognitve) role of the regions that may be disconnected or resected. Functional exploration of cognitive functions (e.g. motricity, language, memory) is therefore essential before any attempt is made to remove the area containing the epileptogenic zone. Defining the epileptogenic network and its functional role involves the transient implantation of electrodes to record the intracranial EEG (iEEG). These intracranial measurements provide valuable information for surgical decision-making (Llorens et al., 2011; Mercier et al., 2022). The implanted electrodes can consist of electrode arrays placed on the surface of the brain (electrocorticography, ECoG) or on probes that penetrate deep brain structures to record stereotactic EEG (sEEG) (see Figure 1 below). In the epileptology department at La Timone, patients are implanted using stereotactic EEG, which allows recording deep structures in the medial temporal cortex, the primary auditory cortex and more in general allows a highly individualized implantation scheme depending on the location of the epileptogenic zone.

Once the electrodes have been implanted, patients remain under continuous medical supervision in the monitoring unit, generally for a period of one to two weeks, depending on the results of the assessments and their clinical condition. This period then offers a unique opportunity to work with patients on fundamental research questions. Firstly, because the electrodes generally cover several cortical regions and subcortical structures in several lobes, sometimes in both cerebral hemispheres ; secondly, because these patients often have near-normal or normal neurological and cognitive functioning, which can enable, to a certain extent,comparison with the healthy human brain. Fundamental research tasks remain limited, as they do not necessarily provide rapid feedback that can be used by clinicians. However, the underlying idea is that these tasks may become one day the ones used in routine by clinicians to map cognitive functions. Indeed, during their hospitalization, patients undergo numerous clinical tests to assess their cognitive functions.

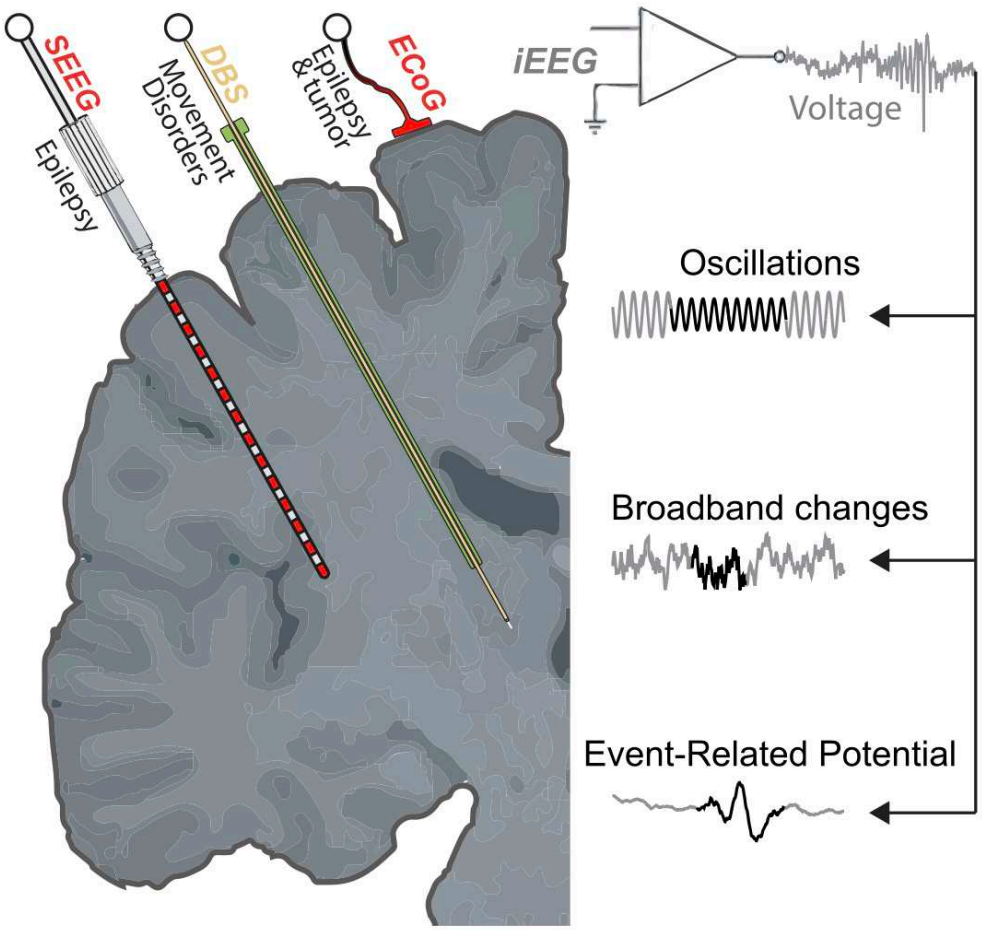


Figure 1 – Intracranial EEG recordings and basic signal features. iEEG can be measured using three different recording strategies: stereotacticEEG (sEEG), Electrocorticogram (ECoG), and deep-brain stimulation (DBS). Signal features are generally characterised as oscillations, broadband changes, andevent-related potentials. *From Mercier et al., 2022.*

2. Clinical functional mapping of cognitive functions

In this chapter, I will only expose the clinical tasks carried out in the epileptology department at La Timone. The final objective is to be able to communicate the regions whose activity is involved in essential cognitive functions, such as language, memory and motricity. In the context of my thesis, I will only describe language tasks and investigations. With this information, neurosurgeons will know which regions must absolutely be preserved and which can be resected. The various clinical tasks (which may also be associated with invasive procedures) are therefore considered to be functional localizers, each allowing a specific cognitive function to be assessed.

a. Pure tones and BaPa

An initial step of the functional mapping is the localization of the auditory cortex. This is done via passive listening to different types of stimuli. Pure tones and the early latency responses they evoke (auditory evoked potentials, AEPs), allow the anatomical segregation of primary and secondary auditory cortex (see Figure 2 below). The primary auditory cortex (PAC) is defined by the presence of primary components (latencies of less than 30 ms) generated in the tip of Heschl's gyrus (HG), which physiologically delimits the primary auditory cortex (broadmann area 41). The secondary auditory cortex (SAC) is defined by late components (greater than 80 ms) generated in broadmann area 42, the lateral parts of the HG and the planum temporale (PT), and in the posterior part of the STG (broadmann area 42) (Trébuchon et al, 2021).

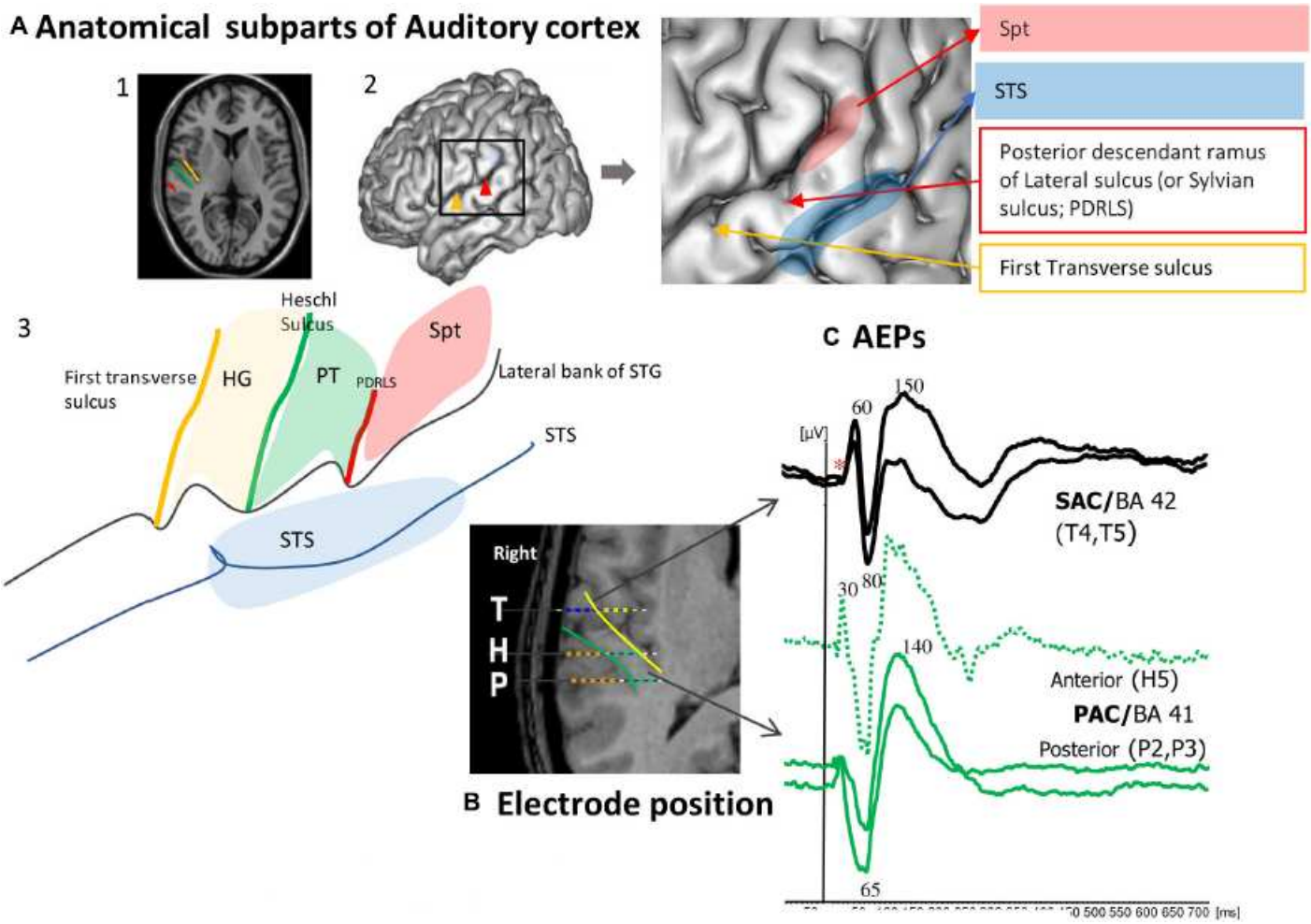


Figure 2 – Functional heterogeneity of auditory cortices. (A) Anatomical subpart of Auditory Cortices. The subparts of auditory cortices are delimited according to the different sulcus of the posterior part of the temporal Gyrus. Heschl's Gyrus (HG) by the transverse sulcus (yellow line) and posteriorly by Heschl sulcus (HS, green line). The Planum Temporale (PT) is limited anteriorly by the HS and posteriorly the horizontal PT is limited where the Sylvian sulcus splits into ascending ramus and descending ramus (red line). The Sylvian-Parieto-temporal region (Spt in red) lies between Posterior descendant ramus of lateral sulcus (PDRLS) and the posterior ascendant ramus of the sylvian sulcus.

The several subparts according to the anatomical landmark are showed on (1) MRI axial view of auditory cortex ; (2) 3D lateral brain representation and (3) on a schematicrepresentation of auditory regions described above. (B) Example of electrode position along the auditory cortex. The 3 electrodes cross the two main sulci. The medial contacts of T electrode (yellow dots), explore the planum polare and the lateral contacts (blue dots) after crossing the transverse sulcus (yellow line) record the anterior part of HG (BA 42). H explore the medial part of HG (green dots, BA 41), then after crossing HS (green line) explore PT. The electrode contacts of P explore the medial part of HG (green dots, BA 41), then after crossing HS (green) explore PT. (C) Example of AEPS in the several sub-part of the auditory cortices in response of tone burst sound. Note the difference of the latency of the first component, around 30 ms for primary auditory cortex (BA 41) and 60 ms for secondary auditory cortex. *Absence of early primary components. *From Trébuchon et al., 2021*.

Passive listening to syllables (/ba/, /pa/) is used to define the lateralization of language at the level of the secondary auditory cortex. For this purpose, we exploit a crucial temporal cue in the perception of speech and language, Voice Onset Time (VOT), which is the main phonetic determinant in the discrimination of voiced - voiceless contrast consonants. VOT refers to the time between a phonetically relevant supra-laryngeal event, such as release and glottal pulsation (Lisker & Abramson, 1970; Abramson & Wahlen, 2017). This time can differ across languages, but in French it is a long positive value (~-110 ms) for voiced stop consonants (/b/, /d/, /g/) and a short positive value (~+20 ms) for voiceless stop consonants (/p/, /t/, /k/). The relevance of using VOT for language lateralization is reflected in the fact that voiced and voiceless consonant-vowel (CV) syllables (in this case /ba/ and /pa/) are differentially and temporally encoded in auditory cortex. The /ba/ syllable elicits an N1/P2 complex followed by a later release component (RC) (~200-250 ms). The /pa/ syllable, on the other hand, elicits a single N1/P2 complex (see Figure 3 below). It is by observing the presence or absence of temporal processing (N1/P2 complex + RC) in auditory structures, although this characteristic is not limited to language, that a lateralisation of language can be suggested (Liégeois-Chauvel et al., 1999; Trébuchon et al., 2005; Trébuchon et al., 2020; Trébuchon et al., 2021).

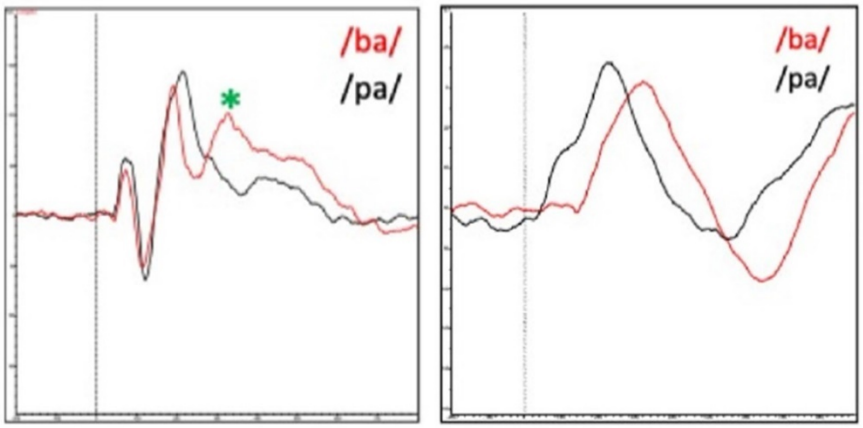


Figure 3 – Temporal course of evoked responses in auditory cortices. Left : A release component (* in green) in response to syllable /ba/ reflecting auditory temporal processing, a biomarker of hemispheric dominance for language. Right : similar ERP patterns in response to syllables /ba/ and /pa/, typically observed in non-dominant hemisphere. *Adapted and rearranged from Trébuchon et al., 2020.*

b. Picture naming

In order to investigate more complex aspects of language, particularly speech production, the picture naming task is proposed to the patients. The experimental material consisted of 36 images of common objects represented in black and white and bearing very consensual names in order to limit possible lexical choices and thus facilitate naming. The objects came from six different semantic categories: accessories, buildings, kitchen utensils, fruit, furniture and musical instruments. The task is organised into six homogeneous blocks, made up of six different images from the same semantic category; and six heterogeneous blocks made up of six images from six different semantic categories. Several lists of blocks are compiled by organising the order of the blocks in a pseudo-random way to vary the alternation between homogeneous and heterogeneous contexts. In addition, within each block, adjacent trials do not contain the same element or elements beginning with the same phoneme. This task involves much more than producing a word : visual recognition of the image, retrieval of the name associated with it at lexical level, and word articulation processes. The high-frequency activity (70-150 Hz) is investigated over epoch periods of 3 seconds (-1 to 2 seconds) relative to the onset of the image on the screen. The analyses used to compute the significant high-frequency activity during the task are carried out using the MIA software (Dubarry et al., 2022). Different regions are likely to respond to this type of task, in particular the temporal structures from 200 ms onwards involved in lexico-semantic processing (activation of the meaning of the image and its possible names) (Llorens et al., 2011; Llorens et al., 2014) but also the prefrontal regions (e.g. inferior frontal gyrus) later on at around 500-600 ms, which are more involved in speech planning and articulation processes (Indefrey & Levelt, 2004). The regions associated with the picture

naming electrodes are therefore considered to be involved in the patient's language network. Alongside investigations of the electrophysiological signal, behavioural results (correct responses, errors) are also taken into account.

c. Electrical stimulation and Wada test

Other investigations, more invasive and coupled with behaviour outcomes, are also used to assess the regions involved in the patient's language network. In particular, electrocortical stimulation is considered to be one of the most direct means to determine whether or not specific functional subunits are involved in a language function, whether in perception or production (Chauvel et al., 1993 ; Trébuchon & Chauvel, 2016). This makes it easier to determine whether structures are likely to lead to language deficits after resection. The stimulation task and mapping strategy are adapted to the exploration of each patient. During electrocortical stimulation, patients are seated in bed and invited to relax. Stimulation is applied from the start of the task and throughout its duration. When performed in temporal auditory regions (primary and/or secondary auditory cortex) during passive listening to sounds, different results can be obtained: auditory hallucinations, i.e. raw, imagined auditory perceptions ; and auditory illusions, i.e. altered perceptions of the sounds heard, such as "The sounds heard seemed louder or clearer, weaker or more distinct, closer or further away" (Trébuchon & Chauvel, 2016 ; Trébuchon et al., 2021). Comprehension deficits can also occur when passively listening to sentences. Electrocortical stimulation is also performed in parallel with speech production tasks such as word repetition, naming, automatic speech (counting) or reading aloud. Various types of deficits can be observed when these tasks are performed: anomia (lack of a word), paraphasia (phonemic, semantic), speech arrest, sensations of numbness. The correlation between a type of error and the area stimulated can thus inform clinicians about the role of these areas in the language network as a whole (Trébuchon & Chauvel ; Trébuchon et al., 2021).

Finally, another test can be used to determine the hemispheric lateralisation of language, the intracarotid amobarbital procedure or Wada test (Wada & Rasmussen, 1960). This method, which was the first to establish hemispheric dominance for language, involves the intracarotid injection of sodium amytal, which induces a temporary loss of function in the ipsilateral cerebral hemisphere. This is a radical and invasive way of finding out whether or not a cerebral hemisphere is involved in the language network, and if it is, the patient will suffer complete aphasia. However, this method is not the first option used and it is only used when the other investigations do not make it possible to define whether functions directly linked to language are present in a given hemisphere. In addition, although the Wada test is very effective because it has very good inter-hemispheric resolution, its intra-hemispheric resolution is very low, which means that it cannot provide precise information about specific regions (Trébuchon et al., 2020).

3. Bridging fundamental research to clinical results

The various clinical tasks used to assess language make it possible to estimate the hemispheric lateralization and functional mapping of the language network for each patient. However, these tasks are not ecological and only assess language in an isolated context. To improve these limitations, fundamental research can contribute by developing tasks involving language in a more ecological context closer to daily language usage. The interactive synchronous speech task developed as part of my thesis work goes in this direction by studying language in a dynamic and interactive context, coupling both the perception and production of language and involving continuous articulatory adjustments (see Chapter 3 - Bridging verbal coordination and neural dynamics).

In the perspective of contributing to clinical functional mapping, this section proposes to compare the results described in chapter 3 with those obtained in the clinical investigation of language in the same patients. The initial step was to define the most relevant clinical results. Following a discussion with the epileptologist responsible of the functional mappin of language, Dr Trebuchon, we decided to focus on two tasks: the passive listening of simple syllables (/ba/, /pa/), a picture naming task and the electrical stimulation (see above). We compared the results of these clinical tasks with the ones obtain for the interactive synchronous speech task i.e. the significant power increases in high-frequency activity (70-125 Hz) and the significant correlations (high-frequency activity power ~ verbal coordination index) (as discussed in Chapter 3). This investigation was conducted at the group level, considering all 16 patients.

a. Comparisons with passive listening (Ba/Pa) and electrical stimulation

The clinical approach consists of comparing the results from various language tasks to suggest a lateralization of language in the brain and to delineate its network with as much precision as possible. In the passive listening task (/ba/, /pa/), the clinician verifies the presence or absence of the N1/P2+RC complex (see section 2.a. Pure tones and Ba/Pa) at the level of the channels located in the primary auditory cortex. The presence of this complex suggests a language lateralization depending on the hemisphere in which the complex is found. This result alone is not clear cut and it needs to be coupled with other results obtained in other language tasks such as picture naming. A significant increase in high-frequency activity (70-150 Hz) during picture naming compared to baseline is used to define the regions potentially involved in the language network. To illustrate the clinical approach, a channel showing a N1/P2+RC complex in the Ba/Pa passive listening task and an increase in high-frequency activity in picture naming will be considered as being highly-likely involved in the language network. The interest of integrating the results of the synchronous speech task is to evaluate to what extent they confirm or differ from the clinical tasks.

Out of the sixteen patients, seven exhibit an N1/P2+RC complex (five in the left hemisphere, two in the right). Within the left hemisphere, among the five patients, twelve channels exhibit the N1/P2+RC complex ; among these twelve channels, eleven respond to both picture naming and interactive synchronous speech, while one channel exclusively responds to interactive synchronous speech. In the right hemisphere, among the patients, two channels exhibit the N1/P2+RC complex, both responding to both picture naming and interactive synchronous speech. Consequently, the results of interactive synchronous speech mirror those of picture naming and offer supplementary insights into a single channel in the left hemisphere.

We conducted similar comparisons using the results from electrical stimulation. Out of the sixteen patients, fourteen exhibit language-related effects (such as anomia, speech arrest, or paraphasia) when electrical stimulation is applied during language production tasks (eleven patients in the left hemisphere and three in the right hemisphere). Within the left hemisphere, among the eleven patients, the electrical stimulation applied on fifty-one channels provoke an effect ; among these channels, twenty-three respond to both picture naming and interactive synchronous speech, eleven respond solely to picture naming, four solely to interactive synchronous speech, and thirteen do not respond to either task. Within the right hemisphere, among the three patients, the electrical stimulation applied on six channels provoke an effect ; among these channels, two respond to both tasks and four respond only to picture naming. Hence, it can be argued that, in the case of electrical stimulation, picture naming is more sensitive across a greater number of channels, making it a better language indicator than interactive synchronous speech. However, it's important to mention that the selection criterion for choosing electrodes for electrical stimulation relies on whether they have previously express a significant response to the picture naming task. The Interactive synchronous speech provides additional insights concerning only four channels in the left hemisphere (otherwise missed by picture naming).

b. Comparisons with picture naming

To assess both the unique information conveyed by interactive synchronous speech and that specific to the picture naming task, and the shared information between these tasks, a group-level comparison similar to the one did in the previous section, was conducted. Out of all the channels in the left hemisphere (1459 channels), 496 show a response to the picture naming. Of these channels, 65% also respond to the interactive synchronous speech and 35% respond only to picture naming. Concerning the right hemisphere, out of all the channels present (936 channels), 300 show a response to the picture naming. Of these channels, 66% also respond to the interactive synchronous speech and 34% to the picture naming only. This initial comparison allowed to assess the extent of the shared responses between the two tasks. The second step was then to determine the location of these common responses and also the location of the responses specific to each task.

Figure 4 illustrates for each hemisphere the channels responding significantly to picture naming and interactive synchronous speech. For both tasks, similar regions are involved, in the temporal, parietal and prefrontal lobes (Figure 4.A). The shared channels between the two tasks confirm these observations, with a strong representation of the superior temporal gyrus bilaterally and the inferior frontal gyrus in the left hemisphere (Figure 4.B). More precisely, channels responding specifically to picture naming and interactive synchronous speech are located in relatively different regions (see Figure 4.C) in both hemispheres. In the left hemisphere, interactive synchronous speech involves more the inferior parietal lobule (IPL BA 40), anteriorsuperior temporal sulcus (aSTS), precuneus and inferior frontal gyrus (IFG BA 44). Picture naming, on the other hand, involves the fusiform gyrus, middle frontal gyrus (MFG BA 9/46), hippocampus and parahippocampal gyrus. The inferior temporal gyrus (ITG BA 20) is involved in both tasks (see Table 1 and Table 2 below). Concerning the right hemisphere, interactive synchronous speech involves the inferior parietal lobule (IPL BA 40), thalamus, insular gyrus and inferior temporal gyrus (ITG BA 20). Picture naming involves the fusiform gyrus and hippocampus, along with the cingulate and orbital gyri. The middle temporal gyrus (BA 21) is also involved in both tasks (see Table 1 and Table 2).

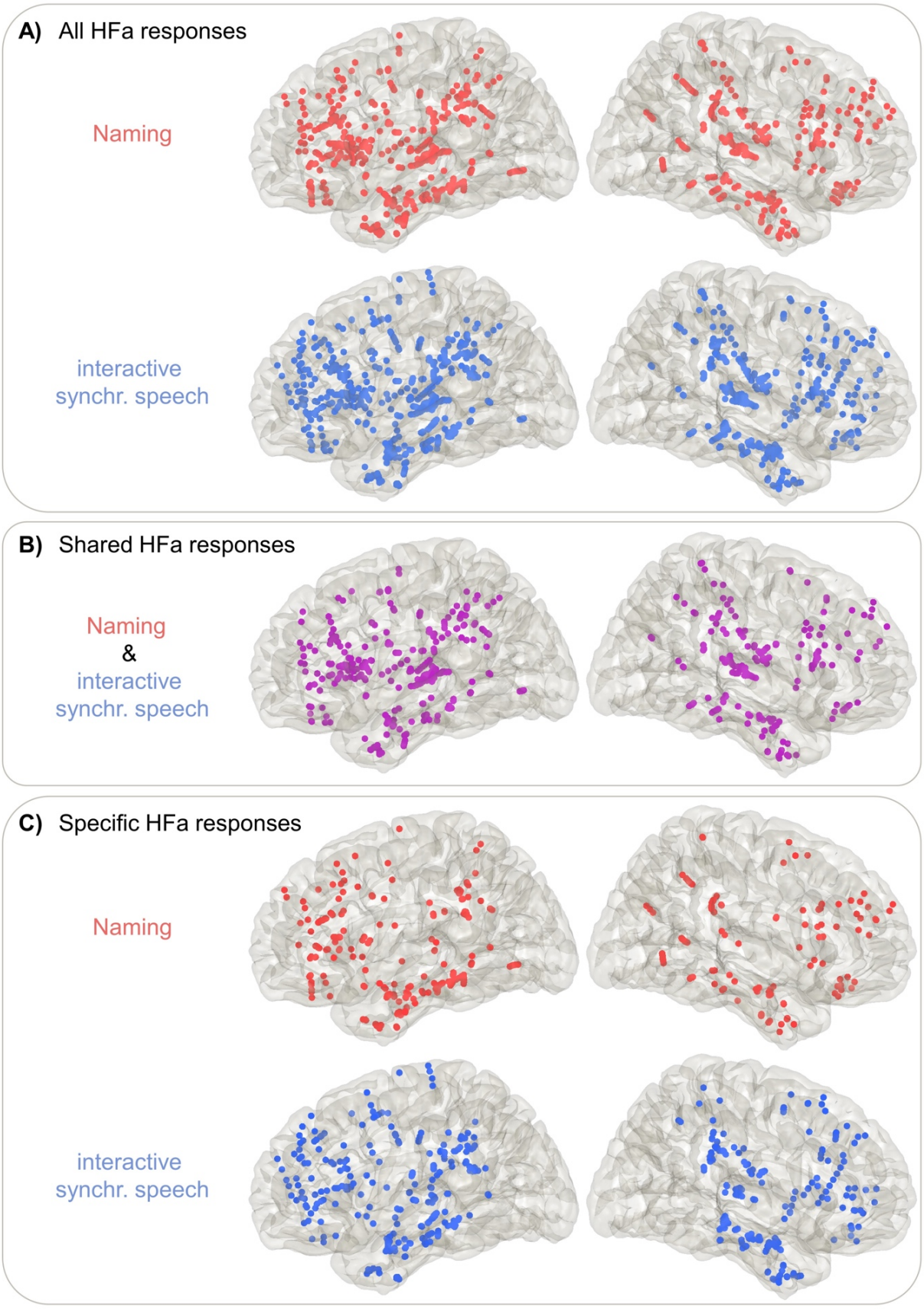


Figure 4 - Illustration of significant HFa responses for picture naming and interactive synchronous speech. Panel (**A**) : representation of all channels showing significant responses to the picture naming (top, red) and interactive synchronous speech (bottom, blue) tasks. Panel (**B**) : representation of common channels to both tasks. Panel (**C**) : representation of channels responding specifically to picture naming (top, red) and interactive synchronous speech (bottom, blue).

Regions (left hemisphere)	occurences
IPL BA 40	26
ITG BA 20	19
MTG aSTS	12
Precuneus	12
IFG BA 44	11
Amygdala	11
MFG BA 9/46	11
MTG BA 21	9
Thalamus	9
STG BA 22	9
IPL BA 39	9
SFG	9
Fusiform gyrus	7
Insular gyrus	7
Precentral gyrus	6
Cingulate gyrus	6
Orbital gyrus	6
Hippocampus	6
MFG BA 6	6
pSTS	5
MFG BA 46	5
Basal ganglia	4
MFG BA 8	4

Regions (right hemisphere)	occurences
ITG BA 20	26
IPL BA 40	17
MTG BA21	16
Thalamus	14
Insular gyrus	13
Fusiform gyrus	10
MFG BA 8	9
MFG BA 9/46	8
Postcentral gyrus	8
IFG BA 44	7
Orbital gyrus	7
MTG aSTS	6
STG TE1/2	5
STG BA 38	4
MFG IFJ	4
SFG	4
MFG BA 6	4
Hippocampus	4

Table 1 – Localization of significant responses specific to the interactive synchronous speech. The two tables (left and right for each hemisphere respectively) show the regions in which the channels respond specifically to interactive synchronous speech task. The 'occurrences' columns indicate the number of channels present in each region. Only regions ≥4 occurrences are displayed.

Regions (left hemisphere)	occurences
Fusiform gyrus	45
ITG BA 20	14
MFG BA 9/46	12
Hippocampus	11
Parahippocampal gyrus	10
Precuneus	9
Insular gyrus	9
Orbital gyrus	9
IFG BA 45	8
IFG BA 44	8
Cingulate gyrus	7
Medioventral occipital cortex	7
IPL BA 40	6
MTG aSTS	5
Precentral gyrus	5
STG BA 38	4
SPL	4
MTG BA 21	4
MFG BA 8	4

Regions (right hemisphere)	occurences
Fusiform gyrus	14
Cingulate gyrus	11
Hippocampus	10
Orbital gyrus	9
MTG BA 21	8
MFG BA 9/46	8
Parahippocampal gyrus	7
Insular gyrus	5
SPL	5
MFG BA 46	4
SFG	4
IPL BA 39	4
IFG BA 44	4

Table 2 – Localization of significant responses specific to the picture naming. The two tables (left and right for each hemisphere respectively) show the regions in which the channels respond specifically to picture naming task. The 'occurrences' columns indicate the number of channels present in each region. Only regions ≥4 occurrences are displayed.

These comparisons uncover regions that are engaged during the two tasks. Specifically, the fusiform gyrus is predominantly activated during picture naming bilaterally, this observation aligns with existing literature linking this region to the semantic processing of visual perception, such as the categorisation of pictures, objects or written words with the visual word form area and the perception of faces with the fusiform face area (Murtha et al, 1999 ; Gauthier et al., 2000 ; McCandliss et al., 2003 ; Kanwisher & Yovel, 2006 ; Mechelli et al., 2006 ; Starrfelt & Gerlach, 2007 ; Price et al., 2010 ; Weiner & Grill-Spector, 2012 ; Forseth et al., 2018). Furthermore, the involvement of the inferior temporal gyrus (ITG) and prefrontal regions (MFG) in response to picture naming in the left hemisphere (see Figure 4.C and Table 2) supports the observations of Riès et al. (2017) suggesting that word retrieval can not be supported by modular view but rather involves a substantial overlap of lexical-semantic activation and word selection mechanisms in the brain. The inferior parietal lobule (IPL BA 40) is predominantly activated bilaterally during interactive synchronous speech. This corresponds with existing literature linking this region to the processing of temporal order in speech syllables, articulation, repetition, motor attention, motor control, and broader motor planning processes (Rushworth et al, 2001 ; Fridriksson et al. 2010 ; Geranmayeh et al., 2012 ; Binkofski et al., 2016 ; Coslett & Schwartz, 2018 ; Zhang et al., 2023). Figure 5

illustrate the five regions (for each hemisphere) with the most occurences for picture naming (Figure 5A) and the interactive synchronoius speech (Figure 5B) in agreement with Table 2 and 1 respectively.

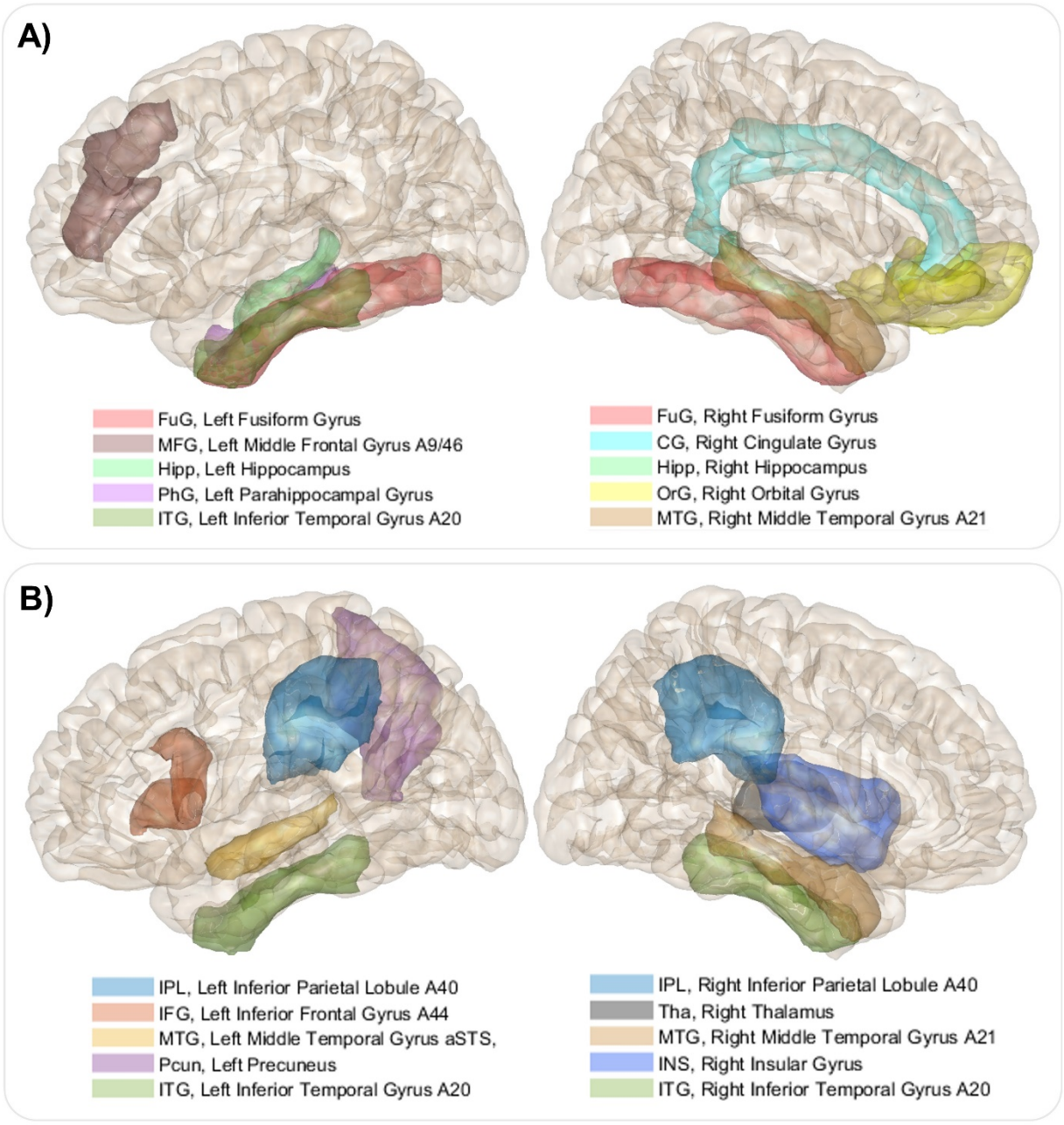


Figure 5 – Illustration of regions with specific responses for picture naming and interactive synchronous speech. A) Regions with specific responses for picture naming, in both hemispheres. **B)** Regions with specific responses for interactive synchronous speech. Legends below each hemisphere indicates the label for each region (according to Brainnetome atlas).

Although preliminary, these comparisons between picture naming and interactive synchronous speech seem to show that the task from my thesis work can provide additional information, particularly concerning regions involved in more complex production processes such as speech planning in a dynamic and interactive context, information that the picture naming task does not provide. It's important to highlight that this comparison is conducted at the group level, encompassing all sixteen patients. Nevertheless, some responses (either for picture naming or interactive synchronous speech) do not show strong consistency across all patients, which makes the approach at group level very difficult. Hence, these comparative investigations only hint at the additional regions that may be highlighted through the utilization of the interactive synchronous speech task. The clinical approach is tailored to each individual patient. Consequently, defining the language network is specific to each patient, taking into account their unique characteristics, including potential functional reorganizations commonly observed in this patient population.

The proposal made in this section is only a first, exploratory step, making it possible to validate the importance of integrating more ecological language tasks into clinical investigations. The objective is not to replace the picture naming task because, as the results show, interactive synchronous speech does not provide entirely the same information as picture naming concerning the language network. These tasks complement each other and a future objective would be to move to the single-patient level to evaluate the benefits of integrating the interactive synchronous speech task while taking into account all the other clinical tasks, the language deficits specific to each patient and also possibly include the results of the surgery both in terms of epilepsy and more in particular associated cognitive disorders.

Chapter 5 – General discussion

This discussion chapter will first summarize the main results obtained during my entire thesis work with a particular focus on the ones obtained in the intracranial study carried out on drug-resistant epileptic patients. Secondly, I will emphasize the fundamental particularity of synchronous speech and how it can contribute to speech motor control models. Finally, I will highlight the limitations and perspectives of my thesis work and an opening point concerning the study of language in neuroscience will be presented.

1. Summary of the results

In my PhD project, we investigated the speech coordinative behaviour using a novel interactive synchronous speech task (Lancia et al., 2017) allowing to study speech perception and production in a dynamic manner. At this aim, we developed a real-time coupled-oscillators virtual partner (VP) that allows – through coupling strength parameters (K) – to modulate the ability to synchronize speech with a speaker. Thus, thanks to the VP's capacity to coordinate and adjust its verbal productions with that of the speaker in real-time, he/she also adapts his/her verbal productions to reach and maintain synchronization. Therefore, there is mutual adaptation between the VP and the speaker. We conducted an initial behavioural study, where participants produced short sentences along with the virtual partner, to assess verbal coordination and the parameters of interest that can influence it. The measure of the verbal coordination index (VCI ; i.e. quantifying the quality of the synchronization) was affected by the values assigned to the strength coupling parameter, making it a key parameter. A low speech rate (3 Hz) generated greater VCI compared to a higher speech rate (4 Hz), and the different stimuli also influenced VCI, favouring the generation of a wide variability in verbal coordination across trials.

The objective was then to investigate the relation between speech coordination and neural dynamics. In this purpose, we adapted the task and leveraged on the excellent spatiotemporal sensitivity of stereotactic electroencephalography (sEEG) recordings from 16 patients with drug-resistant epilepsy while they performed the interactive synchronous speech task. Similarly to the behavioural study, the verbal interaction between patient and VP allowed the creation of a variable coordination context, resulting in a wide range of phase delays across trials. Concerning the neurophysiological correlates, several interesting findings have been underscored. Firstly, the task involving both speech production and perception was efficient to highlight both the dorsal and ventral pathways, from low to high frequency activity. Secondly, spectral profiles of neural responses in the language network were nuanced when coupled with behaviour (verbal coordination index, VCI), highlighting the fact that some regions are « simply » involved in the task, in a general manner, while others are truly sensitive to the quality of verbal coordination. More precisely, four regions of interest showed strong involvement, namely the superior temporal gyrus (STG BA41/42 and BA22), the inferior parietal lobule (IPL BA 40) and the inferior frontal gyrus (IFG BA44). Thirdly, high frequency activity in secondary auditory regions showed a stronger sensitivity to behaviour (VCI) compared to primary auditory regions. Finally, the high frequency activity of the IFG BA44 specifically indexed the online coordinative adjustments that were continuously required to compensate the deviation from synchronization. This latter point emphasized the relation between coordination abilities and anatomical circuitry of the IFG to the neural dynamics of this same region, showing that it plays an important role in the temporal adjustment of speech that are necessary to synchronize to external speech.

Finally, these results were compared with those obtained from clinical investigations. This preliminary exploratory investigation indicated that the results of the interactive synchronous speech task align to a large extent with those observed in both the passive listening task of isolated syllables (Ba/Pa) and the picture naming task. Interestingly, by exploring the location of channels responding significantly to interactive synchronous speech and the location of channels responding significantly to picture naming, regions specifically involved in interactive synchronous speech were defined, providing important additional information for clinicians for the functional mapping of patients' language functions. Overall, these findings underscored the potential to use a fully dynamic, adaptive and interactive language task in the clinical investigation, as it is evaluate a more ecological use of language, i.e. in interaction context.

2. The contribution of synchronous speech to speech motor control models

Speech production requires fine synchronization and coordination of a very large number of muscles with complex mechanical properties in order to set in motion the articulatory structures of the tongue, lips, jaw and larynx in a wide range of configurations (Tremblay et al., 2019). The cascade of motor processes leads to what we define as "speech", which is fundamental for communication. Speech motor control is thus an important area of language research, representing one of the most complex sensorimotor activities we engage in, as we must adapt our productions to the environment and specifically to the interlocutor with whom we are conversing. One of the main questions is how such a complex and flexible system can be controlled to maintain fluent speech.

Several models of speech motor control have been proposed, including the Directions Into Velocities of Articulators (DIVA) model (Tourville & Guenther, 2011 ; Guenther, 2016), the State Feedback Control (SFC; Houde & Nagarajan, 2011), the ACTion-based model of speech production, speech perception and speech acquisition (ACT ; Kröger et al., 2009) and, more recently, the Feedback Aware Control of Tasks in Speech model (FACTS ; Parrell et al., 2019). All these models, with their wide variety of architectures, are characterised by their specific use and their method of combining three essential functional components: feedforward, feedback and predictive model. This implies that speech motor control is based on updating a predictive model. The motor message containing the motor commands to be executed is sent to the motor regions to perform the articulation, and this message is also transmitted to the sensory regions as a prediction message. The prediction model is then updated in proportion to the difference between the actual consequences of the articulation performed (i.e. the actual auditory feedback) and the predictions (i.e. the intended auditory feedback), representing a prediction error signal. By adjusting the motor commands and updating the predictions, speech stability is maintained and any sensory errors are corrected online.

Synchronous speech thus represents an interesting case for speech motor control models, which generally place the processing of self-generated auditory feedback at the center of the orientation of verbal productions (Parrell et al., 2019a ; Parrell et al., 2019b). The many studies on auditory feedback illustrate this well by inducing alterations in auditory feedback and observing compensation phenomena (Houde & Jordan, 1998 ; Lee., 1950 ; Black., 1951 ; Fairbanks., 1955 ; Stuart et al., 2002 ; Yamamoto & Kawabata, 2014 ; Karlin et al., 2021). In contrast, in the case of synchronous speech, the brain simultaneously receives auditory feedback from our own voice and the voice of the partner (speaker or recording). In a similar way, the task developed in my thesis work involves mutual adaptation between a speaker and the virtual partner (VP), meaning that to maintain synchronization of verbal productions we have to adapt them in real time ; the speaker thus perceives his own voice and that of the VP more or less simultaneously according to the quality of his coordination. It is therefore interesting to put synchronous speech into perspective with speech motor control models which do not take into account the influences of external vocal inputs on speech production. The question then arises : how do feedback predictions of one's own speech and that of the speech of others interact during speech production? In the interactive synchronous speech task, the goal is to perform a speech production at the same time as an identical external speech production, so the expectation at the perceptual level is to receive self-generated auditory feedback that is identical to an external speech input. At the neural level, the results obtained in my thesis work show that high-frequency activity in the superior temporal gyrus (specifically STG BA 22) decreases as coordination (or synchronization) increases. Interpreting this as speaker-induced suppression (SIS), which is characterised by a suppression of the neural response corresponding to the speech signal of one's own voice relative to the speech signals of others (Niziolek et al. 2013 ; Kurteff et al., 2023), leads to the suggestion that sensorimotor processes drive speech movements to align with an external target (VP speech signal) rather than an internally defined prediction. The VP's speech signal, occurring simultaneously with his/her own production, is treated as if it were self-generated. Synchronous speech thus addresses a situation that models of speech motor control do not fully take into account. It can be assumed that the feedback predictions of one's own speech and the other's speech converge to represent a single prediction. There is therefore a need for speech motor control models to incorporate the influences of external speech inputs on sensorimotor processes during speech, a point already emphasized by Sato et al. (2012) and Bradshaw & McGettigan (2021). Interactive synchronous speech in this case highlights the ability to change articulatory representations flexibly and dynamically in response to external speech inputs, a process in which the inferior frontal gyrus (IFG BA44) is likely to be involved, its activity having been shown to specifically index the online coordination adjustments that are continually required to compensate for deviation from synchronization.

3. Limitations and future directions

The work presented in this PhD dissertation illustrated the design of a synchronous speech task in an interaction context with a virtual partner able to adapt its verbal productions to those of a speaker in real time. To bridge verbal coordination with neural dynamics, we focused on a population of drug-resistant epileptic patients, a population that provides access to extremely rich intracranial data with excellent temporal resolution and spatial specificity. As with any experimental project, we had to make some critical choices and explored methodological and practical possibilities in order to make the experiments feasible. As a result, a number of limitations can be identified at various levels, as well as prospects for future investigations.

The most challenging constraint was to adapt the task to the clinical context. In order to be carried out by this specific patient population, particularly prone to fatigue because of the intracranial electrode implantation operation, the task had to be short enough to enable as many patients as possible to perform it, while at the same time ensuring that a minimum amount of data was collected to enable consistent analyses to be carried out. In this purpose, the task was set to 24 trials, one trial representing a sequence of repetitions lasting 14 seconds. Depending on the patients' state of fatigue and motivation, we proposed a second session of 24 trials. Of the 16 patients who took part in the study, 6 were able to complete two full sessions. This choice allowed to include as many patients as possible in the somewhat limited time-span of two years. However, to achieve this it was necessary to restrict the number of parameters included in the task, particularly at the level of the virtual partner. Parameters included the speech stimuli articulatory variability and complexity, the speech rate, and the coupling strength (K) inducing variable coordination behaviours in the virtual partner. We also had the possibility to modify the fundamental frequency of the virtual partner's productions, precisely by applying arbitrary increases or decreases, generating high-pitched or low-pitched speech stimuli respectively. This last parameter was discarded to reduce the number of trials in a session and to focus the project exclusively on relating the verbal coordination between the patient and the virtual partner to the electrophysiological signal. This meant to rule out further investigations into the potential convergence phenomena that may occur during interactive speech tasks. Indeed, the study by Bradshaw & McGettigan (2021) showed that such phenomena (interlocutors' speech characteristics becoming more similar to each other's) are indeed present at the fundamental frequency during synchronous speech tasks (see section The case of synchronous speech). It would therefore be very interesting to investigate these aspects as part of a future application of this task and to be able to study the neural correlates of convergence phenomena in a dynamic context. It would also be interesting to assess the extent to which convergence phenomena influence verbal coordination or whether the opposite is true, and even more interestingly whether the coordination behaviour of the virtual partner interacts with these potential phenomena.

A second limitation, which is common to all studies with intracranial recordings, is the difference in the brain regions explored across patients. This is a major difficulty when we want to conduct statistical analyses at group level and particularly considering regions of interest. Although for my thesis work it was possible to conduct analyses by region of interest, it was difficult to gain access to regions such as the premotor or motor regions which are less frequently explored than the temporal regions. Motor regions play an essential and central role in language production, which in the context of a task such as synchronous speech would have been a definite advantage in terms of reconstructing the dorsal language pathway almost in its entirety. In addition, my thesis work focused mainly on the left hemisphere, again as a result of the different regions explored in the patients. It is very difficult to obtain bilaterally explored regions for each patient, which is an additional limitation if we want to study lateralization. The question of hemispheric asymetry was not the biggest limitation in the case of my thesis, as the main language processes, production aspects such as verbal coordination and speech planning, could be investigated to a very large extent in the left hemisphere. All these constraints, which were unknown to me at the start of my thesis, enabled me to approach the research in a different way, forcing me to find new angles of investigation, new ways of approaching analyses and, above all, to redouble my efforts and make the best possible use of the data acquired, however incomplete they may be. Particularly, one way to gauge the strength of observed effects and to address the challenges posed by inter-individual differences in intracranial studies is to examine the consistency of these effects across the population. In this way, the results are stronger if, for a given region, all or a large majority of patients show the same effects, patterns of responses. This was the approach taken in the study presented in the article thesis concerning analysis by regions of interest which is a good compromise between the fixed-effect and the mixed-effect modeling approaches.

A last limitation is the lack in exploiting the dynamic nature of the interactive synchronous speech task. Indeed, the temporal dimension was very early on partly discarded due to the trial-by-trial averaging over time to obtain, on the one hand, a unique verbal coordination indexes (VCI) to each trial and, on the other hand, mean power values in each frequency band of the electrophysiological signal. Although this made it possible to conduct effective correlation analyses to couple our behavioural and neural data, the only analysis exploiting the temporal aspect is the phase-amplitude coupling analysis carried out between the speech signal and the high-frequency activity of the electrophysiological signal. In addition, an important analysis that could provide further information is missing, namely a directional connectivity analysis. Specifically, it would be interesting to investigate the connectivity between the temporal regions (superior temporal gyrus) and the prefrontal regions (inferior frontal gyrus) and to assess the extent to which the quality of coordination (VCI) influences the connectivity between these two regions. One avenue would be to use information theory (Cover & Thomas, 1991), which represents one of the most effective and promising data-driven frameworks for quantifying task-related information in neural signals and has been widely applied to neurophysiological signals (Ince et al., 2017 ; Giordano et al., 2017).

In particular, information-theoretic measures based on the Wiener-Granger principle or also known as Granger causality (Brovelli et al., 2004 ; Seth et al., 2015), such as directed information, mutual information and conditional mutual information (Ince et al., 2012), quantify directional influences at the network level and can capture arbitrary time-shifted conditional dependencies (linear and non-linear) between neural signals (Combrisson et al., 2022). This would make it possible to infer whether activity localised in the inferior frontal gyrus (IFG BA44) that is modulated according to online coordination adjustments exerts a top-down influence (i.e. sends a prediction message) towards the superior temporal gyrus (STG BA22) and how these influences operate as a function of the degree of coordination in real time.

4. About language in neuroscience research

Language is undoubtedly one of the richest and most complex functions in human beings. Hence, understanding its neural mechanisms is, if possible, an extremely complex challenge. Historically, the first approach was to study the perception and production of speech and language separately. Then, the community realized the importance of the relationship between perception and production, gradually enriching the theoretical models of language. But what is the essence of language if not its social aspect? This is why studying language in isolation is not enough if we want to understand more broadly how the brain processes language. In interaction there is a need to coordinate speaking and listening, to integrate one's own (verbal) actions with the actions of the interlocutor and to adapt language flexibly to the interlocutor and the social context. The context of interaction was an essential point in my thesis work. How can we study language in a context of interaction while investigating brain activity? This is the question that is becoming increasingly important in the neuroscience of language. Because being in interaction means not being able to maintain a level of experimental control equivalent to that which we can afford to have in an isolated context. This raises questions about the choice of paradigms, their ecological validity and also the technique to record the brain signal. To gain further understanding of the relation between these elements , I illustrate below some of paradigms from the studies previously described in Chapter 1 (see section II. The language in interaction) and some other language paradigms. This representation, that does not pretend to cover the ensemble of all language tasks, has the benefit of allowing to evaluate the relation between the ecological validity of the tasks while considering the time scale on which they take place, and hence the types of analyses they allow.

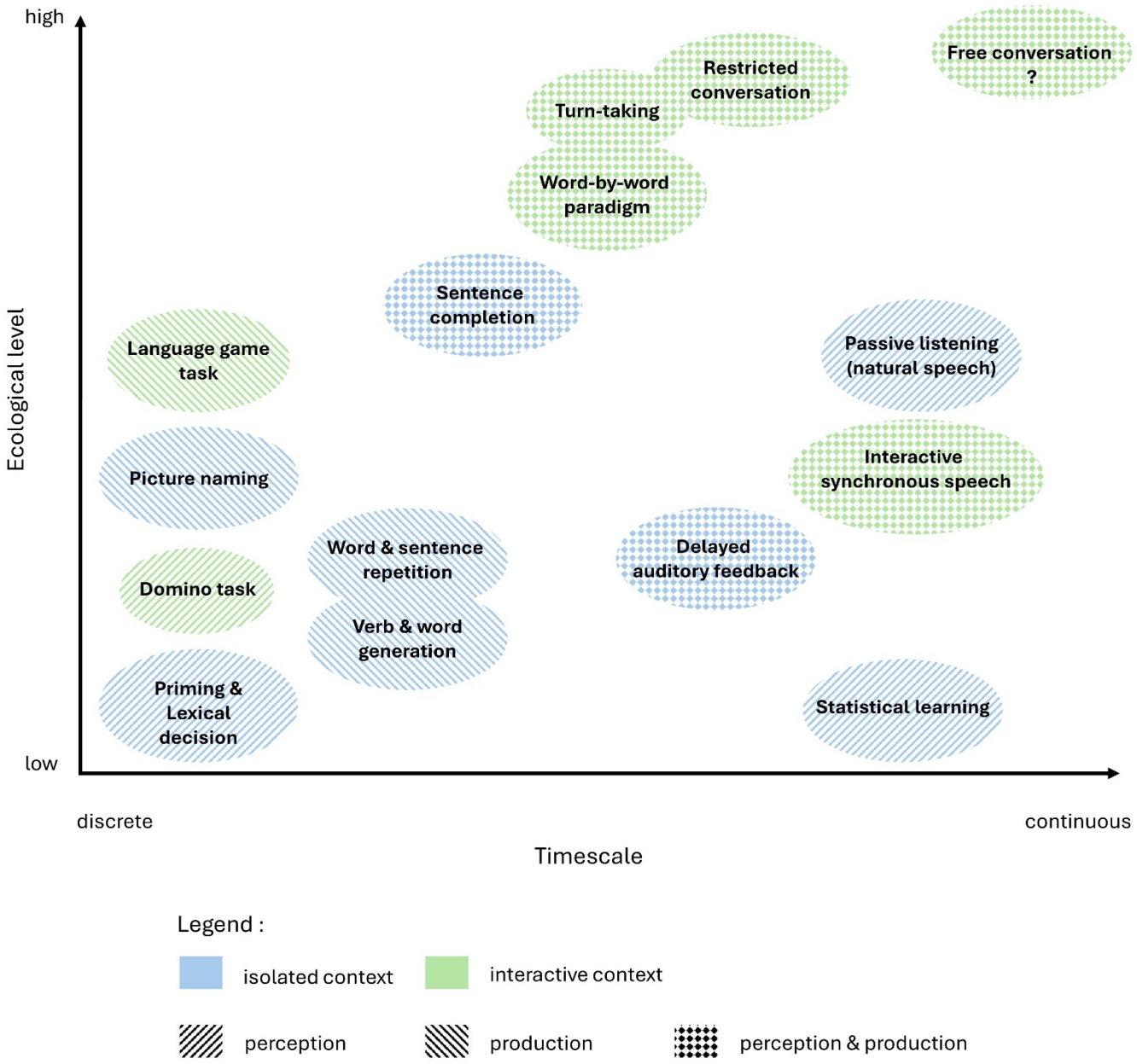


Figure 1 – Illustration of language study paradigms as a function of the temporal scale of events/analyses and the ecological validity. The x-axis represents the timescale over which the task is performed (discrete to continuous). The y-axis represents the ecological validity of the task (low to high), the higher it is, the closer the task is to a natural/real situation. Blue and green colors define the context in which the language is studied, isolated or interactive respectively. The motifs (stripes and checks) indicate whether the paradigm in question allow to investigate speech perception, production or both.

Different tasks can be used depending on the phenomena we want to study, and each task can be characterised according to its ecological validity, i.e. how close it is to a real life situation. The context (isolated or interactive) in which the participants perform the tasks also influences this ecological validity. As a first example, the interactive synchronous speech task developed during my thesis work presents a somewhat low level of ecology insofar as this behaviour is limited to particular contexts (e.g. churches and stadium). However, it requires sustained verbal productions for several seconds (14 seconds) with a virtual partner adapting its verbal productions in real time. There is therefore an interactive, dynamic and continuous use of language. It is a task that is carried out in interaction and not in an isolated context, although not with another person.

This figure offers different reading levels and interpretation concerning the way in which language is studied and enables paradigms to be put into perspective in relation to others. Different tasks can be used depending on the phenomena we want to study, and each task can be characterised according to its level of ecology. On the one hand, this level of ecology can be characterised by the stimuli used, a point which echoes the criteria proposed by Hamilton & Huth (2020) concerning the three questions on which the naturalness of a stimulus is based : Firstly, is it a stimulus to which a person could reasonably be exposed outside an experimental setting? Secondly, does the stimulus appear in the same context as in real life? Thirdly, is the participant's motivation to perceive and understand the stimulus specific to the experimental setting or is it a motivation that the participant would develop in real life? On the other hand, the level of ecology can be characterised by the context in which the participants perform the task. On this point, the vast majority of experimental tasks are performed in recording rooms or boxes. The use of techniques such as functional magnetic resonance imaging (fMRI) or magnetoencephalography (MEG) reduces the ecological framework even further because the participant is almost completely immobilized. This figure can be considered from several angles, with the ecological validity axis concerning both the stimuli used and the context in which they are perceived/used, providing even more information and classifying the tasks and paradigms more precisely.

In a similar way, paradigms can be organised according to the timescale on which they take place, i.e. the extent to which a task is organised in discrete or continuous events. For example, passive natural language listening tasks (Simony et al., 2016; Huth et al., 2016; de Heer et al., 2017; Hamilton et al., 2018; te Rietmolen et al., 2023) expose participants to excerpts of pre-recorded stories or film clips that can last several minutes without interruption. In contrast, picture naming tasks (Indefrey & Levelt, 2004) involve more discrete events, organised into multiple picture naming. And in parallel with the timescale of the tasks, we can consider the possible analyses: can we analyze the brain signal continuously or discretely? Can the signal be analyzed using time-continuous analysis techniques such as the temporal response function (TRF ; Crosse et al., 2016 ; Brodbeck et al., 2023), or is it better to analyze the signal using evoked related potentials (ERPs)? A distinction can be made between the temporal structure of a task and the temporal window of the analyses we carry out. An example of this is conversation. A

conversation takes place continuously, with the roles of speaker and listener alternating, but when we look at the analyses that are mostly carried out, we see that the studies are interested in the phenomenon of turn-taking, reducing the time windows for analyses to discrete portions.

At a first look, the interactive synchronous speech task developed during my thesis work presents a somewhat low to moderate level of ecology insofar as this behaviour can be only mostly observed in particular contexts (e.g. churches and stadium). However, it requires sustained verbal productions for several seconds (14 seconds) with a virtual partner adapting his verbal productions in real time. There is therefore a dynamic and continuous use of language in an interactive context. It was therefore possible to exploit long time windows for our analyses. In addition, this is a task that is carried out in interaction and not in an isolated context, although not with another person. A task can be considered as "interactive" when the participant uses language with another person or an artificial agent, with the intention to communicate or respond. The domino task (Mukherjee et al., 2018), the language game task (Boux et al., 2021), the word-by-word paradigm (Fjaellingsdal et al., 2020), the turn-taking tasks (Bögels et al., 2015 ; Bögels et al., 2018 ; Bögels et al., 2020) and the restricted conversation (Pérez et al., 2017) are all interactive tasks. In the domino task participants take turns producing a word alternatively in accordance with a rhyming rule (e.g. 'outside', 'sidewalk'). The language game task involves role-playing by asking a participant to name an object placed in front of him/her in two different contexts, one social and the other isolated. In the word-by-word paradigm participants to co-construct sentences by producing a word in turn. The turn-taking tasks correspond to question-and-answer exchanges between a participant and the experimenter, similar to the exchanges we might have in daily life. In the restricted conversation task participants, in the same room but without seeing each other, discuss a topical subject (e.g. sport, travel, cinema) over a set number of turns.

There is thus a wide variety of tasks and paradigms that allow the interactional aspect of language to be emphasized, but this requires a number of constraints to be taken into account. The higher the ecological validity of a task, the more difficult it is to control experimentally. How do we define and isolate what we want to study? It is in response to this question that turn-taking periods, for instance, have been the target of numerous investigations.These periods are essential components of verbal interactions such as conversation, and make it possible to study the phenomena of motor preparation of speech and anticipation of turn-taking. Similarly, sentence completion tasks (Magyari et al., 2017 ; Piai et al., 2017 ; Gastaldon et al., 2020) also investigate these phenomena. Language is more than just speaking and listening. This line of research therefore allows to study predictive aspects of speech production and perception over periods that are not subject to the constraints imposed by neurophysiological signal recording techniques. Indeed, it is difficult to record a clean neurophysiological signal during speech production as it is likely to be contaminated by electro-myographic artefacts generated by the articulatory muscles (Goncharova et al., 2003 ; Vos et al., 2010 ; Piai et al., 2015 ; Jiang et al., 2019). Electroencephalography (EEG) is particularly sensitive to these artefacts, and it is therefore very difficult, if not sometimes impossible, to analyze

the EEG signal during natural speech production. One of the solutions lies in intracranial recordings, which are not subject to electro-myographic artefacts and have excellent temporal and spectral resolution. This makes it possible to analyze the neurophysiological signal in a continuous manner as instance using delayed auditory feedback tasks, allowing to further enrich our understanding of speech production processes and motor control mechanisms (Ozerk et al., 2022). Similarly, it is thanks to intracranial recordings that it has been possible to investigate neural dynamics in a task such as interactive synchronous speech involving sustained and continuous speech production. The signal recording technique could also represent a dimension to be added to the Figure 1, allowing to distinguish the possible paradigms that can be conducted according to each technique.

An important point to consider regarding the different recording techniques is the case of hyperscanning, which involves recording the brain activity of two or more individuals simultaneously while they perform a task. This methodology, which makes extensive use of functional near-infrared spectroscopy (fNIRS) and EEG, is aimed at investigating the mechanisms involved in social cognition and particularly at studying similarities between the brain activities of individuals. Thus, it is not just language that is being targeted, but large areas of investigation such as decision making, problem solving, knowledge sharing, joint attention, affective communication, emotional regulation and musical practice (Liu et al., 2018 ; Redcay & Schilbach, 2019 ; Czeszumski et al., 2020 ; Kelsen et al., 2022). The experimental paradigms developed for these research topics thus involve individuals in realistic, structured, ecologically valid and dynamically unfolding interactions. However, when it comes to brain data, several criticisms can be raised. For instance, quantifying the co-variation of neural dynamics between two brains is difficult to interpret as it is likely to reflect shared cognitive processing of external environmental stimuli that are presented to both individuals (Hamilton et al., 2021). To a certain extent, then, it is interesting to ask: does recording two individuals in an interaction really teach us more than if we had only recorded one?

Finally, a last point of reflection corresponds to what we did during my thesis work, i.e. investigating neurophysiological data as a function of a behavioural measure. Using the interactive synchronous speech task, an index of verbal coordination was calculated and correlated with the power acitivity in different frequency bands. In this way we were able to specify a verbal coordination activity network different from that obtained by the classic task-baseline contrast. This approach is interesting since it allows the neurophysiological data to be explained from a new angle and thus offers the possibility of using different behavioural indexes to couple them with brain activity and thus characterize activity networks linked to behavioural performance. This approach was previously used in my Master's thesis to study the relationship between prediction and comprehension in the context of conversation (Hidalgo et al., 2022 ; see Appendix). In this EEG study, participants watched a series of videos showing two interlocutors conversing. We parametrically degraded speech, measured eye movements, EEG and speech comprehension in order to investigate the mechanisms underlying speech comprehension under challenging conditions. We built on the fact that participants

change their gaze according to the interlocutor and used this measure as a prediction indicator to classify dialogue turns as more or less well predicted and to study evoked potentials accordingly. In this way, electrophysiological data were coupled with behavioural data, offering a coherent analysis perspective directly linked to the conversational context. This way of coupling the behaviour with electrophysiological data could also be considered as an additional dimension to Figure 1, adding further information on the extent of data exploitation.

All this raises to the central question posed by the Figure 1 : What can we actually conclude when using a specific task and analysis pipeline ? Although the representation is far from being exhaustive, it nevertheless provides a classification framework to resituate and contextualize the research conducted on language and, most importantly, enables us to pinpoint the areas that are under-represented and those that have yet to be explored. The study of conversation is becoming increasingly prominent in neuroscience research, and it is important to ask how it can be studied. Based on the axes shown in Figure 1, the ideal would be to design a completely ecological paradigm, i.e. free natural conversation, allowing the signal to be analyzed over extended time windows. However, is it possible to design such a paradigm? And if so, is it necessary to study all the processes taking place during the conversation? It is undeniable that the context of interaction is essential to consider, but the diversity of the paradigms developed is a richness per se, and each provides important information that contributes to a better understanding of how the brain processes language.

A final thought would be to add a dimension on the number of studies carried out for each paradigm, and look at Figure 1 from a different perspective (see Figure 2 below), bringing to the foreground paradigms widely investigated, and we would see that these ones would mainly be paradigms with a low ecological validity but representing important and pioneering topics of language research, such as priming and lexical decision (Bentin et al, 1985 ; Praamstra et al., 1993 ; Neely, 2012) or statistical learning (Saffran et al., 1996 ; Romberg & Saffran, 2010). Whereas in the background, « at the horizon », we would distinguish paradigms involving language in the context of interaction, less explored.

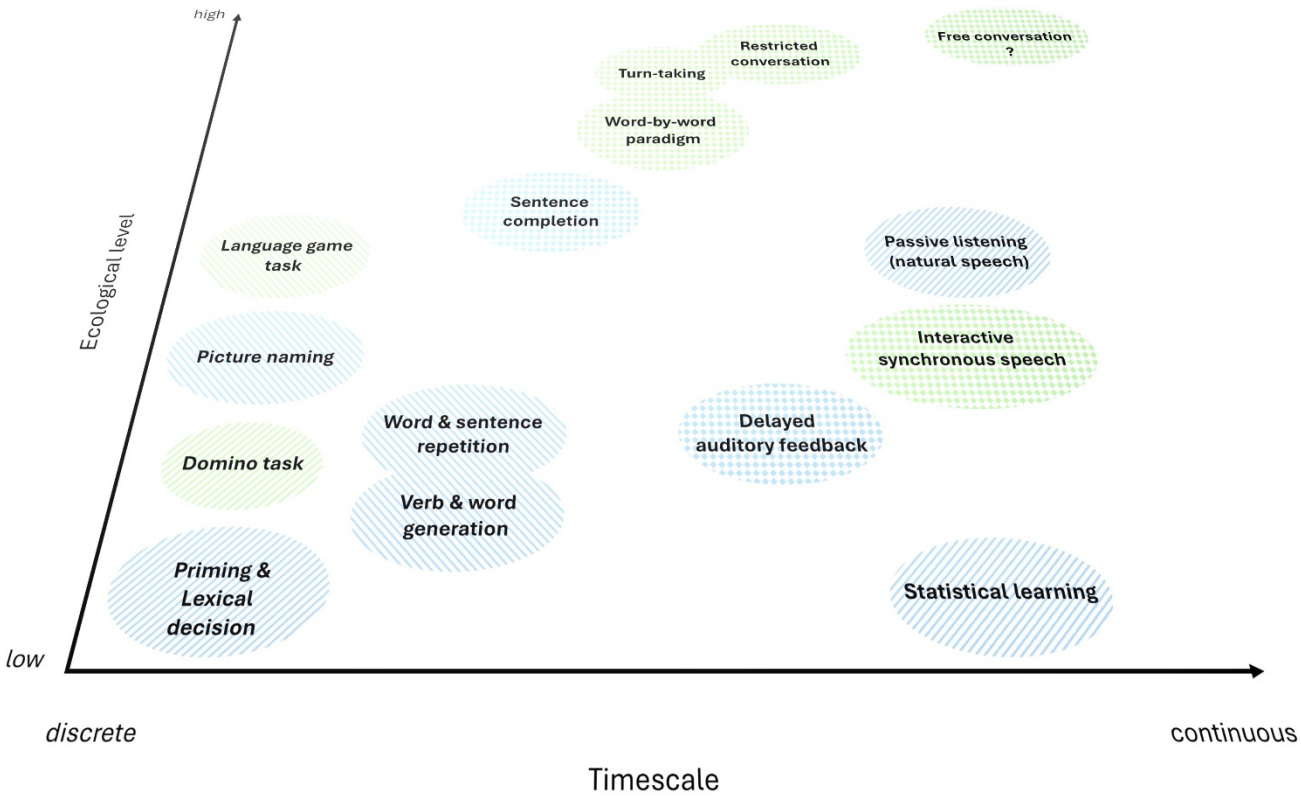


Figure 2 – Projection of Figure 1. The depth illustrates the representativeness of paradigms in neuroscience research on language. In the forefront, the main research topics that have been widely explored. In the background, the less explored research topics.

Beyond illustrating and highlighting what has been done, this offers a prism of perspectives, fields to explore, a waybill. There is still a long way to go in these directions, and improvements in techniques, particularly for recording brain signals, offer such possibilities that it is important to take advantage of and exploit in the best possible way. It's a horizon that the neuroscience community need to venture towards.

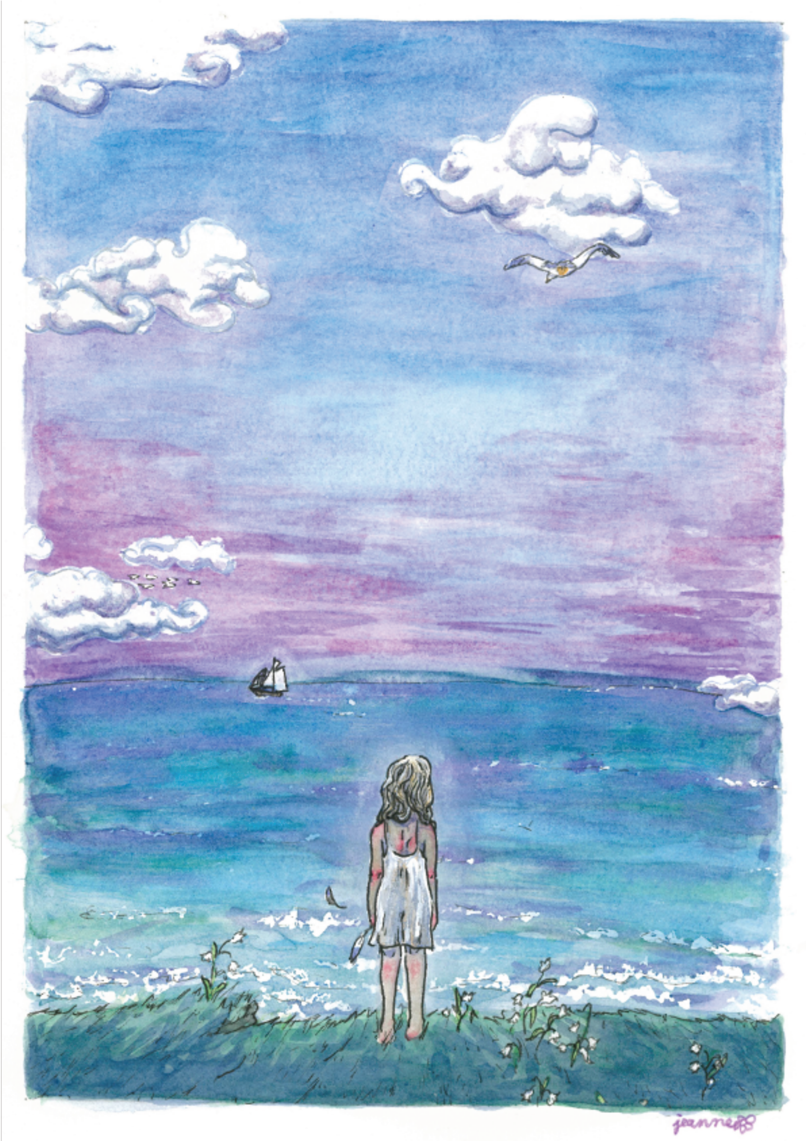


Illustration by Jeanne Bayle from the book “les petites variations.”
by Isaïh Schwab-Mohamed.

Conclusion

As my PhD comes to an end, I think it's important to conclude by addressing a few essential points. These almost four years represent my first real experience in scientific research. It was long, dense, rich and challenging, sometimes desperate but also stimulating. From my Master's research internship onwards, I was lucky enough to join an incredible team, a team that I stayed with for my PhD. And despite a troubled context when I started, I appreciated what the dynamics of scientific research were like right from the start, surrounded by warm, passionate and funny PhD students.

As for many students before me, I experienced the frustration of not being able to correctly run my analysis pipeline, of feeling that things weren't moving forward and of seeing time pass before my eyes. But I've also learnt what it feels like when (all) the problems are solved, when step by step our project takes shape, and finally when we realize that sometimes, often, we're the only ones who can solve our project's problems. Because that's one of the most important things I've learnt from my PhD : research is a succession of problems that we face every day and that we have to solve in order to (maybe) have an answer to the question we're asking ourselves. These are stages that are repeated, each time a bit different. And each time, we have to be a bit less wrong.

With that in mind, one of the particularities of my project involved working with intracranial data, which made it difficult both in terms of the time needed to obtain enough data and to analyze it. But before collecting data, I first met patients. The task developed in my PhD was not an easy one, but it gave me the opportunity to accompany each of the patients who carried it out. And it was through all these acquisition sessions that I realized even more how important it is for research in general to respond first and foremost to the needs of society. Since patients are not concerned with trendy theories or the latest analyses that can be performed on their brain activity, the only thing they want is to be able to get better. This only reflects my personal experience, but it's an aspect of research that I feel is still under-emphasized. We must not forget why we do research and, above all, for whom.

Bibliography

Abramson, A. S., & Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. Journal of phonetics, 63, 75-86.

Alario, F. X., Chainay, H., Lehericy, S., & Cohen, L. (2006). The role of the supplementary motor area (SMA) in word production. Brain research, 1076(1), 129-143.

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., & Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature neuroscience*, *22*(4), 627-632.

Bailey, L. M., Bodner, G. E., Matheson, H. E., Stewart, B. M., Roddick, K., O'Neil, K., ... & Fawcett, J. M. (2021). Neural correlates of the production effect: An fMRI study. Brain and Cognition, 152, 105757.

Basilakos, A., Smith, K. G., Fillmore, P., Fridriksson, J., & Fedorenko, E. (2018). Functional characterization of the human speech articulation network. Cerebral Cortex, 28(5), 1816-1830.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical Neurophysiology*, *60*(4), 343-355.

Bernstein, L. (1967). Treatment of velopharyngeal incompetence. Archives of Otolaryngology, 85(1), 67-74.

Bhaya-Grossman, I., & Chang, E. F. (2022). Speech computations of the human superior temporal gyrus. Annual review of psychology, 73, 79-102.

Binkofski, F. C., Klann, J., & Caspers, S. (2016). On the neuroanatomy and functional role of the inferior parietal lobule and intraparietal sulcus. In Neurobiology of language (pp. 35-47). Academic Press.

Black, J. W. (1951). The effect of delayed side-tone upon vocal rate and intensity. Journal of Speech and Hearing Disorders, 16(1), 56-60.

Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. Cognition, 203, 104347.

Bögels, S., & Levinson, S. C. (2017). The brain behind the response: Insights into turn-taking in conversation from neuroimaging. Research on Language and Social Interaction, 50(1), 71-89.

Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. Neuropsychologia, 109, 295-310.

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no... How the brain interprets the pregnant pause in conversation. PloS one, 10(12), e0145474.

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2020). Conversational expectations get revised as response latencies unfold. Language, Cognition and Neuroscience, 35(6), 766-779.

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

Boux, I., Tomasello, R., Grisoni, L., & Pulvermüller, F. (2021). Brain signatures predict communicative function of speech production in interaction. cortex, 135, 127-145.

Bradshaw, A. R., & McGettigan, C. (2021). Convergence in voice fundamental frequency during synchronous speech. PloS one, 16(10), e0258747.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. Cognition, 75(2), B13-B25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of experimental psychology: Learning, memory, and cognition, 22(6), 1482.

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). Bulletin et Memoires de la Societe anatomique de Paris, 6, 330-357.

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., ... & Simon, J. Z. (2023). Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions. *Elife*, *12*, e85012.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., & Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. Proceedings of the National Academy of Sciences, 101(26), 9849-9854.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. Phonology, 6(2), 201-251.

Bürki, A., & Laganaro, M. (2014). Tracking the time course of multi-word noun phrase production with ERPs or on when (and why) cat is faster than the big cat. Frontiers in psychology, 5, 79843.

Cappa, S. F., & Perani, D. (2003). The neural correlates of noun and verb processing. Journal of Neurolinguistics, 16(2-3), 183-189.

Castellucci, G. A., Kovach, C. K., Howard III, M. A., Greenlee, J. D., & Long, M. A. (2022). A speech planning network for interactive language use. Nature, 602(7895), 117-122.

Chan, A. M., Dykstra, A. R., Jayaram, V., Leonard, M. K., Travis, K. E., Gygi, B., ... & Cash, S. S. (2014). Speech-specific tuning of neurons in human superior temporal gyrus. Cerebral Cortex, 24(10), 2679-2693.

Chartier, J., Anumanchipalli, G. K., Johnson, K., & Chang, E. F. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron, 98(5), 1042-1054.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. Journal of personality and social psychology, 76(6), 893.

Chauvel, P., Landré, E., Trottier, S., Vignel, J. P., Biraben, A., Devaux, B., & Bancaud, J. (1993). Electrical stimulation with intracerebral electrodes to evoke seizures. Advances in neurology, 63, 115-121.

Christoffels, I. K., Formisano, E., & Schiller, N. O. (2007). Neural correlates of verbal feedback processing: an fMRI study employing overt speech. Human brain mapping, 28(9), 868-879.

Chung, J. E., Sellers, K. K., Leonard, M. K., Gwilliams, L., Xu, D., Dougherty, M. E., ... & Chang, E. F. (2022). High-density single-unit human cortical recordings using the Neuropixels probe. Neuron, 110(15), 2409-2421.

Clark, H. H. (1996). Using language. Cambridge university press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22(1), 1-39.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production : Evidence from the priming of noun-phrase structure. Journal of Memory and Language, 49(2), 214-230.

Combrisson, E., Allegra, M., Basanisi, R., Ince, R. A., Giordano, B. L., Bastin, J., & Brovelli, A. (2022). Group-level inference of information-based measures for the analyses of cognitive brain networks from neurophysiological data. NeuroImage, 258, 119347.

Coslett, H. B., & Schwartz, M. F. (2018). The parietal lobe and language. Handbook of Clinical Neurology, 151, 365-375.

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. Proceedings of the National Academy of Sciences, 106(50), 21442-21446.

Cover, T. M. (1999). Elements of information theory. John Wiley & Sons.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, *10*, 604.

Cummins, F. (2002). On synchronous speech. Acoustics Research Letters Online, 3(1), 7-11.

Cummins, F. (2003). Practice and performance in speech produced synchronously. Journal of Phonetics, 31(2), 139-148.

Cummins, F. (2004). Synchronization among speakers reduces macroscopic temporal variability. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 26, No. 26).

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. Journal of Phonetics, 37(1), 16-28.

Czeszumski, A., Eustergerling, S., Lang, A., Menrath, D., Gerstenberger, M., Schuberth, S., ... & König, P. (2020). Hyperscanning: a valid method to study neural inter-brain underpinnings of social interaction. Frontiers in Human Neuroscience, 14, 39.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. Journal of Neuroscience, 37(27), 6539-6557.

De Jaegher, H., Peräkylä, A., & Stevanovic, M. (2016). The co-creation of meaningful action: bridging enaction and interactional sociology. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1693), 20150378.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. Journal of Neuroscience, 39(39), 7722-7736.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. In Advances in experimental social psychology (Vol. 33, pp. 1-40). Academic Press.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nature neuroscience, 19(1), 158-164.

Dixon, S. (2005, August). An On-Line Time Warping Algorithm for Tracking Musical Performances. In IJCAI (pp. 1727-1728).

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. The Journal of the Acoustical Society of America, 95(5), 2670-2680.

Dubarry, A. S., Liégeois-Chauvel, C., Trébuchon, A., Bénar, C., & Alario, F. X. (2022). An open-source toolbox for Multi-patient Intracranial EEG Analysis (MIA). NeuroImage, 257, 119251.

Dumas, G., de Guzman, G. C., Tognoli, E., & Kelso, J. S. (2014). The human dynamic clamp as a paradigm for social interaction. Proceedings of the National Academy of Sciences, 111(35), E3726-E3734.

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010). Inter-brain synchronization during social interaction. PloS one, 5(8), e12166.

Duncan, S., 1972. Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23, 283–292.

Fairbanks, G. (1955). Selective vocal effects of delayed auditory feedback. Journal of speech and hearing disorders, 20(4), 333-346.

Flinker, A., Chang, E. F., Kirsch, H. E., Barbaro, N. M., Crone, N. E., & Knight, R. T. (2010). Single-trial speech suppression of auditory cortex activity in humans. Journal of Neuroscience, 30(49), 16643-16650.

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., & Crone, N. E. (2015). Redefining the role of Broca's area in speech. Proceedings of the National Academy of Sciences, 112(9), 2871-2875.

Foerster, O. (1931). The cerebral cortex in man. Lancet, 2, 309-312.

Formisano, E., Kim, D. S., Di Salle, F., Van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron, 40(4), 859-869.

Forseth, K. J., Kadipasaoglu, C. M., Conner, C. R., Hickok, G., Knight, R. T., & Tandon, N. (2018). A lexical semantic hub for heteromodal naming in middle fusiform gyrus. Brain, 141(7), 2112-2126.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. Language production, 1, 373-420.

Fridriksson, J., Kjartansson, O., Morgan, P. S., Hjaltason, H., Magnusdottir, S., Bonilha, L., & Rorden, C. (2010). Impaired speech repetition and left parietal lobe damage. Journal of Neuroscience, 30(33), 11057-11061.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. Physiological reviews.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nature reviews neuroscience, 11(2), 127-138.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue : A study in conceptual and semantic co-ordination. Cognition, 27(2), 181-218.

Gastaldon, S., Arcara, G., Navarrete, E., & Peressotti, F. (2020). Commonalities in alpha and beta neural desynchronizations during prediction in language comprehension and production. *Cortex*, *133*, 328-345.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform “face area” is part of a network that processes faces at the individual level. Journal of cognitive neuroscience, 12(3), 495-504.

Geranmayeh, F., Brownsett, S. L., Leech, R., Beckmann, C. F., Woodhead, Z., & Wise, R. J. (2012). The contribution of the inferior parietal cortex to spoken language production. Brain and language, 121(1), 47-57.

Geschwind, N. (1970). The Organization of Language and the Brain: Language disorders after brain damage help in elucidating the neural basis of verbal behavior. Science, 170(3961), 940-944.

Giles, H., Coupland, N., & Coupland, I. U. S. T. I. N. E. (1991). 1. Accommodation theory: Communication, context, and. Contexts of accommodation: Developments in applied sociolinguistics, 1.

Giordano, B. L., Ince, R. A., Gross, J., Schyns, P. G., Panzeri, S., & Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. elife, 6, e24763.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nature neuroscience, 15(4), 511-517.

Golfinopoulos, E., Tourville, J. A., & Guenther, F. H. (2010). The integration of large-scale neural network modeling and functional brain imaging in speech motor control. Neuroimage, 52(3), 862-874.

Goncharova, I. I., McFarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2003). EMG contamination of EEG: spectral and topographical characteristics. *Clinical neurophysiology*, *114*(9), 1580-1593.

Goregliad Fjaellingsdal, T., Schwenke, D., Scherbaum, S., Kuhlen, A. K., Bögels, S., Meekes, J., & Bleichner, M. G. (2020). Expectancy effects in the EEG during joint and spontaneous word-by-word sentence production in German. Scientific Reports, 10(1), 5460.

Gracco, V. L., & Lofqvist, A. (1994). Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements. Journal of Neuroscience, 14(11), 6585-6597.

Grasso-Cladera, A., Costa-Cordella, S., Mattoli-Sanchez, J., Vilina, E., Santander, V., Hiltner, S. E., & Parada, F. J. (2024). Embodied Hyperscanning for studying social interaction: A scoping review of simultaneous brain and body measurements.
Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. Computer Speech & Language, 25(3), 601-634.

Guenther, F. H. (2016). Neural control of speech. Mit Press.

Gwilliams, L., King, J. R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. Nature communications, 13(1), 6606.
Gwilliams, L., Marantz, A., Poeppel, D., & King, J. R. (2023). Top-down information shapes lexical processing when listening to continuous speech. Language, Cognition and Neuroscience, 1-14.

Hakim, U., De Felice, S., Pinti, P., Zhang, X., Noah, J. A., Ono, Y., ... & Tachtsidis, I. (2023). Quantification of inter-brain coupling: A review of current methods used in haemodynamic and electrophysiological hyperscanning studies. NeuroImage, 120354.

Hamilton, A. F. D. C. (2021). Hyperscanning: beyond the hype. *Neuron*, *109*(3), 404-407.

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. Language, cognition and neuroscience, 35(5), 573-582.

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Current Biology, 28(12), 1860-1871.

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. Cell, 184(18), 4626-4639.

Hashimoto, Y., & Sakai, K. L. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. Human brain mapping, 20(1), 22-28.

Hasson, U. (2016, July). Face to face, brain to brain: exploring the mechanisms of dyadic social interactions. In International Journal of Psychology (Vol. 51, pp. 873-874). 2-4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXON, ENGLAND: ROUTLEDGE JOURNALS, TAYLOR & FRANCIS LTD.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends in cognitive sciences, 16(2), 114-121.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. Current directions in psychological science, 2(3), 96-100.

Heilbron, M., & Chait, M. (2018). Great expectations: is there evidence for predictive coding in auditory cortex? Neuroscience, 389, 54-73.

Hickok, G. (2012). Computational neuroanatomy of speech production. Nature reviews neuroscience, 13(2), 135-145.

Hillis, A. E., Rorden, C., & Fridriksson, J. (2017). Brain regions essential for word comprehension: Drawing inferences from patients. Annals of neurology, 81(6), 759.

Hirano, S., Kojima, H., Naito, Y., Honjo, I., Kamoto, Y., Okazawa, H., ... & Konishi, J. (1997). Cortical processing mechanism for vocalization with auditory verbal feedback. Neuroreport, 8(9), 2379-2382.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. Science, 279(5354), 1213-1216.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. Frontiers in human neuroscience, 5, 82.

Howard III, M. A., Volkov, I. O., Abbas, P. J., Damasio, H., Ollendieck, M. C., & Granner, M. A. (1996). A chronic microelectrode investigation of the tonotopic organization of human auditory cortex. Brain research, 724(2), 260-264.

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

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453-458.

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A. L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. elife, 4, e06213.

Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. Human brain mapping, 38(3), 1541-1573.

Ince, R. A., Mazzoni, A., Bartels, A., Logothetis, N. K., & Panzeri, S. (2012). A novel test to determine the significance of neural selectivity to single and multiple potentially correlated stimulus features. Journal of neuroscience methods, 210(1), 49-65.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. Frontiers in psychology, 2, 255.

Indefrey, P., & Levelt, W. J. (2000). The neural correlates of language production. The new cognitive neurosciences; 2nd ed., 845-865.

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

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience*, *9*(4), 304-313.

Jiang, X., Bian, G. B., & Tian, Z. (2019). Removal of artifacts from EEG signals: a review. Sensors, 19(5), 987.

Johnson, J. F., Belyk, M., Schwartze, M., Pinheiro, A. P., & Kotz, S. A. (2019). The role of the cerebellum in adaptation: ALE meta-analyses on sensory feedback error. *Human brain mapping*, *40*(13), 3966-3981.

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., ... & Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. Nature, 551(7679), 232-236.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. Philosophical Transactions of the Royal Society B: Biological Sciences, 361(1476), 2109-2128.

Karlin, R., Naber, C., & Parrell, B. (2021). Auditory feedback is used for adaptation and compensation in speech timing. Journal of Speech, Language, and Hearing Research, 64(9), 3361-3381.

Keller, P. E., Novembre, G., & Hove, M. J. (2014). Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1658), 20130394.

Kelsen, B. A., Sumich, A., Kasabov, N., Liang, S. H., & Wang, G. Y. (2022). What has social neuroscience learned from hyperscanning studies of spoken communication? A systematic review. Neuroscience & Biobehavioral Reviews, 132, 1249-1262.

Kelso, J. S. (2008). Haken-kelso-bunz model. Scholarpedia, 3(10), 1612.

Kerbrat-Orecchioni, C. (1998). L'interaction épistolaire. La lettre entre réel et fiction. Paris : SEDES, 15-36.

Khader, P. H., Jost, K., Mertens, M., Bien, S., & Rösler, F. (2010). Neural correlates of generating visual nouns and motor verbs in a minimal phrase context. Brain Research, 1318, 122-132.

Khanna, A. R., Muñoz, W., Kim, Y. J., Kfir, Y., Paulk, A. C., Jamali, M., ... & Williams, Z. M. (2024). Single-neuronal elements of speech production in humans. Nature, 626(7999), 603-610.

Klein, G. (2001). Features of team coordination. New trends in cooperative activities: Understanding system dynamics in complex environments, 68-95.

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. Organizational simulation, 53, 139-184.

Kotz, S. A., & Schmidt-Kassow, M. (2015). Basal ganglia contribution to rule expectancy and temporal predictability in speech. *Cortex*, *68*, 48-60.

Kotz, S. A., Stockert, A., & Schwartze, M. (2014). Cerebellum, temporal predictability and the updating of a mental model. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130403.

Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. Speech Communication, 51(9), 793-809.

Kuhlen, A. K., & Abdel Rahman, R. (2023). Beyond speaking: neurocognitive perspectives on language production in social interaction. Philosophical Transactions of the Royal Society B, 378(1875), 20210483.

Kuramoto, Y. (1975). Self-entrainment of a population of coupled non-linear oscillators. In International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan (pp. 420-422). Springer Berlin Heidelberg.

Kurteff, G. L., Lester-Smith, R. A., Martinez, A., Currens, N., Holder, J., Villarreal, C., ... & Hamilton, L. S. (2023). Speaker-induced suppression in EEG during a naturalistic reading and listening task. Journal of Cognitive Neuroscience, 35(10), 1538-1556.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. Journal of statistical software, 82(13).

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. Journal of nonverbal behavior, 27, 145-162.

Lakretz, Y., Ossmy, O., Friedmann, N., Mukamel, R., & Fried, I. (2021). Single-cell activity in human STG during perception of phonemes is organized according to manner of articulation. NeuroImage, 226, 117499.

Lancia, L., Chaminade, T., Nguyen, N., & Prevot, L. (2017, August). Studying the link between inter-speaker coordination and speech imitation through human-machine interactions. In Interspeech 2017 (pp. 859-863).

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. Journal of speech and hearing research, 14(4), 677-709.

Lee, B. S. (1950). Effects of delayed speech feedback. The Journal of the Acoustical Society of America, 22(6), 824-826.

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., ... & Chang, E. F. (2024). Large-scale single-neuron speech sound encoding across the depth of human cortex. Nature, 626(7999), 593-602.

Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A. Y., Tal, I., ... & Schroeder, C. E. (2020). Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. Science advances, 6(33), eabb0977.

Levelt, W. J. (1983). Monitoring and self-repair in speech. Cognition, 14(1), 41-104.

Levelt, W. J. (1999). Models of word production. Trends in cognitive sciences, 3(6), 223-232.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. Trends in cognitive sciences, 20(1), 6-14.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. Frontiers in psychology, 6, 136034.

Liégeois-Chauvel, C., De Graaf, J. B., Laguitton, V., & Chauvel, P. (1999). Specialization of left auditory cortex for speech perception in man depends on temporal coding. Cerebral cortex, 9(5), 484-496.

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the 6th international congress of phonetic sciences (Vol. 563, pp. 563-567).

Liu, D., Liu, S., Liu, X., Zhang, C., Li, A., Jin, C., ... & Zhang, X. (2018). Interactive brain activity: review and progress on EEG-based hyperscanning in social interactions. Frontiers in psychology, 9, 1862.

Lizarazu, M., Lallier, M., & Molinaro, N. (2019). Phase– amplitude coupling between theta and gamma oscillations adapts to speech rate. Annals of the new York Academy of Sciences, 1453(1), 140-152.

Llorens, A., Trébuchon, A., Liégeois-Chauvel, C., & Alario, F. X. (2011). Intra-cranial recordings of brain activity during language production. Frontiers in psychology, 2, 375.

Llorens, A., Trébuchon, A., Riès, S., Liégeois-Chauvel, C., & Alario, F. X. (2014). How familiarization and repetition modulate the picture naming network. Brain and language, 133, 47-58.

Long, M. A., Katlowitz, K. A., Svirsky, M. A., Clary, R. C., Byun, T. M., Majaj, N., ... & Greenlee, J. D. (2016). Functional segregation of cortical regions underlying speech timing and articulation. Neuron, 89(6), 1187-1193.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. Cognitive science, 36(8), 1404-1426.

Luo, J., Hage, S. R., & Moss, C. F. (2018). The Lombard effect: from acoustics to neural mechanisms. Trends in neurosciences, 41(12), 938-949.

Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology*, *8*, 230845.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. Journal of Experimental Psychology : General, 127(4), 331.

Martikainen, M. H., Kaneko, K. I., & Hari, R. (2005). Suppressed responses to self-triggered sounds in the human auditory cortex. Cerebral cortex, 15(3), 299-302.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. Trends in cognitive sciences, 7(7), 293-299.

Mechelli, A., Sartori, G., Orlandi, P., & Price, C. J. (2006). Semantic relevance explains category effects in medial fusiform gyri. Neuroimage, 30(3), 992-1002.

Mercier, M. R., Dubarry, A. S., Tadel, F., Avanzini, P., Axmacher, N., Cellier, D., ... & Oostenveld, R. (2022). Advances in human intracranial electroencephalography research, guidelines and good practices. Neuroimage, 260, 119438.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174), 1006-1010.

Mirman, D., Kraft, A. E., Harvey, D. Y., Brecher, A. R., & Schwartz, M. F. (2019). Mapping articulatory and grammatical subcomponents of fluency deficits in post-stroke aphasia. Cognitive, Affective, & Behavioral Neuroscience, 19, 1286-1298.

Moerel, M., De Martino, F., & Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. Journal of Neuroscience, 32(41), 14205-14216.

Montague, P. R., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., ... & Fisher, R. E. (2002). Hyperscanning: simultaneous fMRI during linked social interactions. Neuroimage, 16(4), 1159-1164.

Morillon, B., Liegeois-Chauvel, C., Arnal, L. H., Bénar, C. G., & Giraud, A. L. (2012). Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. Frontiers in psychology, 3, 248.

Mukherjee, S., Badino, L., Hilt, P. M., Tomassini, A., Inuggi, A., Fadiga, L., ... & d'Ausilio, A. (2019). The neural oscillatory markers of phonetic convergence during verbal interaction. Human brain mapping, 40(1), 187-201.

Murtha, S., Chertkow, H., Beauregard, M., & Evans, A. (1999). The neural substrate of picture naming. Journal of cognitive neuroscience, 11(4), 399-423.

Neely, J. H. (2012). Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes in reading*, 264-336.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. Journal of Neuroscience, 33(41), 16110-16116.

Nourski, K. V. (2017). Auditory processing in the human cortex: An intracranial electrophysiology perspective. Laryngoscope investigative otolaryngology, 2(4), 147-156.

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. Trends in cognitive sciences, 23(11), 913-926.

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. Science advances, 5(11), eaay6279.

Oxenham, A. J., Boucher, J. E., & Kreft, H. A. (2017). Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy. The Journal of the Acoustical Society of America, 142(3), EL264-EL269.

Ozker, M., Doyle, W., Devinsky, O., & Flinker, A. (2022). A cortical network processes auditory error signals during human speech production to maintain fluency. PLoS biology, 20(2), e3001493.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America, 119(4), 2382-2393.

Parrell, B., & Houde, J. (2019). Modeling the role of sensory feedback in speech motor control and learning. Journal of Speech, Language, and Hearing Research, 62(8S), 2963-2985.

Parrell, B., Lammert, A. C., Ciccarelli, G., & Quatieri, T. F. (2019). Current models of speech motor control: A control-theoretic overview of architectures and properties. The Journal of the Acoustical Society of America, 145(3), 1456-1481.

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. PLoS computational biology, 15(9), e1007321.

Paulk, A. C., Kfir, Y., Khanna, A. R., Mustroph, M. L., Trautmann, E. M., Soper, D. J., ... & Cash, S. S. (2022). Large-scale neural recordings with single neuron resolution using Neuropixels probes in human cortex. Nature Neuroscience, 25(2), 252-263.

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

Penfield, W., & Roberts, L. (1959). Speech and brain mechanisms. Princeton University Press.

Pérez, A., Carreiras, M., & Duñabeitia, J. A. (2017). Brain-to-brain entrainment: EEG interbrain synchronization while speaking and listening. Scientific reports, 7(1), 4190.

Piai, V., Meyer, L., Dronkers, N. F., & Knight, R. T. (2017). Neuroplasticity of language in left-hemisphere stroke: Evidence linking subsecond electrophysiology and structural connections. *Human brain mapping*, *38*(6), 3151-3162.

Piai, V., Riès, S. K., & Knight, R. T. (2015). The electrophysiology of language production: what could be improved. *Frontiers in Psychology*, *5*, 121297.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27, 169–225.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and brain sciences, 36(4), 329-347.

Pickles, J. O. (2015). Auditory pathways: anatomy and physiology. Handbook of clinical neurology, 129, 3-25.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. Nature reviews neuroscience, 21(6), 322-334.

Praamstra, P., & Stegeman, D. F. (1993). Phonological effects on the auditory N400 event-related brain potential. *Cognitive Brain Research*, *1*(2), 73-86.

Price, C. J., Noppeney, U., Phillips, J., & Devlin, J. T. (2003). How is the fusiform gyrus related to category-specificity? Cognitive Neuropsychology, 20(3-6), 561-574.

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. Nature Reviews Neuroscience, 20(8), 495-505.

Riès, S. K., Dhillon, R. K., Clarke, A., King-Stephens, D., Laxer, K. D., Weber, P. B., ... & Knight, R. T. (2017). Spatiotemporal dynamics of word retrieval in speech production revealed by cortical high-frequency band activity. *Proceedings of the National Academy of Sciences*, *114*(23), E4530-E4538.

Riès, S. K., Pinet, S., Nozari, N. B., & Knight, R. T. (2021). Characterizing multi-word speech production using event-related potentials. Psychophysiology, 58(5), e13788.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906-914.

Rushworth, M. F., Krams, M., & Passingham, R. E. (2001). The attentional role of the left parietal cortex: the distinct lateralization and localization of motor attention in the human brain. Journal of cognitive neuroscience, 13(5), 698-710.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. In Studies in the organization of conversational interaction (pp. 7-55). Academic Press.

Saenz, M., & Langers, D. R. (2014). Tonotopic mapping of human auditory cortex. Hearing research, 307, 42-52.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., & Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. Science, 326(5951), 445-449.

Salmelin, R., Hari, R., Lounasmaa, O. V., & Sams, M. (1994). Dynamics of brain activation during picture naming. Nature, 368(6470), 463-465.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J. L., & Nguyen, N. (2013). Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production. Frontiers in psychology, 4, 51019.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. Analyzing discourse: Text and talk, 71, 71-93.

Schuhmann, T., Schiller, N. O., Goebel, R., & Sack, A. T. (2009). The temporal characteristics of functional activation in Broca's area during overt picture naming. cortex, 45(9), 1111-1116.

Schwartze, M., & Kotz, S. A. (2013). A dual-pathway neural architecture for specific temporal prediction. *Neuroscience & Biobehavioral Reviews*, *37*(10), 2587-2596.

Seth, A. K., Barrett, A. B., & Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. Journal of Neuroscience, 35(8), 3293-3297.

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology : Human Perception and Performance, 29(2), 326.

Silveri, M. C. (2021). Contribution of the cerebellum and the basal ganglia to language production: Speech, word fluency, and sentence construction—evidence from pathology. *The Cerebellum*, *20*(2), 282-294.

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. Nature communications, 7(1), 12141.

Sörös, P., Cornelissen, K., Laine, M., & Salmelin, R. (2003). Naming actions and objects: cortical dynamics in healthy adults and in an anomic patient with a dissociation in action/object naming. Neuroimage, 19(4), 1787-1801.

Starrfelt, R., & Gerlach, C. (2007). The visual what for area: words and pictures in the left fusiform gyrus. Neuroimage, 35(1), 334-342.

Steinschneider, M., Fishman, Y. I., & Arezzo, J. C. (2008). Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. Cerebral Cortex, 18(3), 610-625.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, 106(26), 10587-10592.

Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. The Journal of the Acoustical Society of America, 111(5), 2237-2241.

Takaso, H., Eisner, F., Wise, R. J., & Scott, S. K. (2010). The effect of delayed auditory feedback on activity in the temporal lobe while speaking: a positron emission tomography study.

TenHouten, W., Schussel, L., Gritsch, M. F., & Kaplan, C. D. (2023). Hyperscanning and the future of neurosociology. Sociological Methodology, 53(1), 139-157.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. Language and cognitive processes, 26(7), 952-981.

Trébuchon-Da Fonseca, A., Giraud, K., Badier, J. M., Chauvel, P., & Liégeois-Chauvel, C. (2005). Hemispheric lateralization of voice onset time (VOT) comparison between depth and scalp EEG recordings. Neuroimage, 27(1), 1-14.

Trébuchon, A., & Chauvel, P. (2016). Electrical stimulation for seizure induction and functional mapping in stereoelectroencephalography. Journal of Clinical Neurophysiology, 33(6), 511-521.

Trébuchon, A., Alario, F. X., & Liégeois-Chauvel, C. (2021). Functional topography of auditory areas derived from the combination of electrophysiological recordings and cortical electrical stimulation. Frontiers in Human Neuroscience, 15, 702773.

Trébuchon, A., Liegeois-Chauvel, C., Gonzalez-Martinez, J. A., & Alario, F. X. (2020). Contributions of electrophysiology for identifying cortical language systems in patients with epilepsy. Epilepsy & Behavior, 112, 107407.

Tremblay, P., Deschamps, I., & Dick, A. S. (2019). Neuromotor organization of speech production.

Volfart, A., McMahon, K. L., Howard, D., & de Zubicaray, G. I. (2022). Neural correlates of naturally occurring speech errors during picture naming in healthy participants. Journal of Cognitive Neuroscience, 35(1), 111-127.

Vos, D. M., Riès, S., Vanderperren, K., Vanrumste, B., Alario, F. X., Huffel, V. S., & Burle, B. (2010). Removal of muscle artifacts from EEG recordings of spoken language production. Neuroinformatics, 8, 135-150.

Wada, J., & Rasmussen, T. (1960). Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance: experimental and clinical observations. Journal of neurosurgery, 17(2), 266-282.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. Journal of pragmatics, 32(8), 1177-1207.

Weiner, K. S., & Grill-Spector, K. (2012). The improbable simplicity of the fusiform face area. Trends in cognitive sciences, 16(5), 251-254.

Wernicke, C. (1874). Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis. Cohn & Weigert.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. Journal of memory and language, 31(2), 183-194.

Yamamoto, K., & Kawabata, H. (2014). Adaptation to delayed auditory feedback induces the temporal recalibration effect in both speech perception and production. Experimental brain research, 232, 3707-3718.

Yates, A. J. (1963). Delayed auditory feedback. Psychological bulletin, 60(3), 213.

Zhang, W., Yang, F., & Tian, X. (2023). Functional connectivity between parietal and temporal lobes mediates internal forward models during speech production. Brain and Language, 240, 105266.

Appendix – The effect of speech degradation on the ability to track and predict turn structure in conversation

The present appendix contains the study I worked on during my Master's thesis. A study published during the first year of my PhD thesis.

CORTEX 151 (2022) 105–115

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

The effect of speech degradation on the ability to track and predict turn structure in conversation

Céline Hidalgo [a,*], Isaïh Mohamed [a], Christelle Zielinski [b] and Daniele Schön [a]

[a] Aix Marseille Univ, Inserm, INS, Inst Neurosci Syst, Marseille, France
[b] Aix-Marseille Univ, Institute of Language, Communication and the Brain, France

ARTICLE INFO

Article history:
Received 29 June 2021
Reviewed 16 September 2021
Revised 15 November 2021
Accepted 20 January 2022
Action editor Alessandro Tavano
Published online 11 March 2022

Keywords:
Conversation
ERPs
Eye-tracking
Prediction
Turn-taking

ABSTRACT

Conversation represents a considerable amount of the daily language usage and plays an important role in language acquisition. In conversation, listeners simultaneously process their interlocutor's turn and prepare their own next turn. As such the turn-taking dynamics heavily relies on prediction. In other words, listeners avail prior knowledge to constrain both speech perception and production. Here we explored the relation between prediction and comprehension while watching two interlocutors having a conversation. We capitalize on gaze switch as a proxy of predictive behaviour to class dialogue turns as more or less well predicted and explore how this affects dialogue comprehension. Moreover, we study the extent to which speech degradation, by increasing the global uncertainty of the context, affects the relation between predictions, brain correlates of prediction errors (N400) and global comprehension. Results show that 1) listeners direct gaze to the current speaker, in particular in challenging conditions, 2) gaze behaviour possibly affects the semantic processing of the upcoming turn (N400), 3) participants with a more efficient gaze predictive behaviour better solve semantic uncertainties at the turn onset, in particular in the most challenging listening condition. Our findings contribute to a better understanding of the relation between predictions, neural predictions errors and speech comprehension under challenging conditions.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Speech perception is not the result of a real-time decoding of audio-visual information. It is a dynamic process of an anticipatory nature building on 1) the use of the linguistic and extra-linguistic context to generate hypotheses about the upcoming signal, i.e. we make predictions about what we are likely to hear and 2) a comparison of our predictions against what is actually perceived. Thus, when we listen to a conversation between two people, we anticipate, for

* *Corresponding author*. Institut de Neurosciences des Systèmes, Faculté de Médecine, bd Jean Moulin, 13005 Marseille, France.
E-mail address: celine.hidalgo@univ-amu.fr (C. Hidalgo).
https://doi.org/10.1016/j.cortex.2022.01.020
0010-9452/© 2022 Elsevier Ltd. All rights reserved.

example, the words that the speakers will produce according to the linguistic context that precedes their utterances. The more we know about this context, the fewer prediction errors we make and the faster and more accurate our understanding of their conversation. According to predictive coding theory (Clark, 2013), predictions would indeed be a default brain mode of operation for processing sensory information (Friston, 2005). To optimize speech perception, the brain would pre-activate representations of the expected speech input (Molinaro et al., 2016) at a phonological, lexical, semantic or even syntactic level (Pickering & Gambi, 2018). With increasingly accurate predictions, the brain would only need to compute the difference between sensory input and prediction, thereby decreasing the cost of perceptual processing and speeding up the understanding of the spoken message. A well-known fact in the comprehension literature is that the amount of contextual information, i.e. the increase in the precision of prediction, positively influences the speed of word recognition (Tyler & Wessels, 1983). This has been thoroughly studied using event-related potentials, by observing changes in the amplitude of a negative component. Indeed, words with low cloze probability engender a prediction error that is visible in an increase of the N400 amplitude. By contrast the N400 amplitude is reduced when perceiving speech in highly predictive contexts (see for a review Kutas & Federmeier, 2011). However, until now, no study had shown a direct link between contextual predictions and the integration of these predictions in the semantic comprehension process in a conversational context. This is due to the fact that, while prediction errors engender observable brain responses (e.g., N400 or Mismtach Negativity), this is not the case of predictions. Recently, Grisoni and colleagues (Grisoni et al., 2021) described an electrophysiological index of predictions called the "Semantic Prediction Potential", and showed, by measuring the N400 brain response during the perception of more or less contextually constrained sentences, the direct effect of predictions (SPP) on the processing of prediction errors (N400).

Another possibility to infer the ongoing predictions is to use behavioural measures. In the case of language, several studies have used for instance eye movements during reading (Rayner, 1978; Staub, 2015). Similarly, in speech perception, some authors relied on the implicit anticipatory gaze behaviour when listening to different people having a conversation (Casillas & Frank, 2017; Keitel et al., 2013). More precisely, listeners anticipate the end of the turn of the current speaker and switch gaze to the following speaker, before the actual onset of the new turn. This predictive behaviour is particularly important considering that the silence separating speech turns of conversing speakers only lasts a few hundred milliseconds (Levinson, 2016; Stivers et al., 2009). In considering turn-taking in conversation, predictions build on several linguistic cues such as prosody (Roberts et al., 2015), lexical-semantic cues (De Ruiter et al., 2006), syntactic structures (Selting, 1996) or pragmatic cues (Beňuš et al., 2011). When listening to a dialogue (as a third person), a similar behaviour possibly takes place via the generation of internal models, allowing a listener to anticipate in a precise temporal window the speakers' turn-taking.

The aim of the current study is to explore the links between predictions at the turn level, measured by gaze switch, and both local and global semantic processing when listening to a dialogue, as indexed by N400 to turn onset and comprehension score, respectively. Importantly we are interested in studying how these relations evolve when adding uncertainty to the context by degrading the input signal. Indeed, degrading the acoustic signal or adding noise or competing speakers, adds uncertainty and challenges speech comprehension (Mattys et al., 2012; Peelle, 2018). To this aim, we asked participants to watch videos of two people having a conversation manipulating different levels of speech degradation. We measured eye movements to estimate predictive gaze behaviour and used this measure to class turns as well predicted (early gaze switch) or poorly predicted (late gaze switch). We also measured EEG and used the N400 as a marker of prediction error at the turn onset and studied the extent to which this response is influenced by gaze behaviour. Finally, all participants responded to several open questions at the end of each dialogue, allowing estimating a global comprehension measure. Most importantly, we explored the relation between these different variables as a function of speech degradation and increasing uncertainty. We hypothesized that poorly predicted turns (late gaze switch) would engender a larger N400 response compared to well anticipated turns (early gaze switch). Importantly, this effect may change as a function of speech degradation and should be reduced at high speech degradation levels when uncertainty is highest. We also hypothesized that, if predictions are fully implicit and based on the understanding of the dialogue, then gaze switches should behave similarly to comprehension score and decrease from normal to moderately degraded and even more for highly degraded speech.

2. Materials and methods

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedures or analyses was preregistered prior to the research being conducted.

2.1. *Participants*

Forty-eight participants (27 females) were tested after being informed of the procedure of the study, which was approved by the Sud Méditerranée Ethics Committee (ID RCB: 2015-A01490-49). In absence of a known effect size, the sample size was chosen as rather large with respect to N400 studies (Šoškić et al., 2021). Inclusion and exclusion criteria: Participants were between 18 and 62 years old (mean = 32, SD = 6 years). They were native French speakers. The had normal or corrected to normal vision and normal hearing. years); they had no history of speech disorder or neurological disease based on self-report. Prior the experiment, the hearing thresholds were controlled using a 5 dB-step custom screening hearing test made in Python 3.7.9 (Expyriment version .10.0). Data from nine participants were excluded

because of noisy EEG recordings (high impedances or excessive eye/movement artifacts) or track loss for long temporal windows in the eye-tracking dataset, resulting in 39 participants in the final dataset.

2.2. Stimuli

We presented 6 audiovisual stimuli of ~4.5 min each, showing a man and a woman having a conversation. The themes of the conversations differed and were joyful to maintain attention during the whole duration of the dialogue. Dialogues contained on average 59 turns (min = 46; max = 66; see details in table 1). Turns lasted on average 3.75s (SD = 2.75) and global speech rate was approximately 5 Hz. 60% of these turns were questions/responses adjacency pairs (min = 56%; max = 66%) in order to elicit predictive behaviours as suggested in Casillas & Frank, 2017 (Casillas & Frank, 2017). We used Final Cut Pro X to set the gap duration at speakers' turn to 500 ms in order to have homogeneous turn conditions, allow anticipatory gaze behaviour (cf. Casillas & Frank, 2017; Foulsham et al., 2010) and obtain equivalent EEG baseline for turn onset analyses.

For the clarity of the evoked potentials, each turn onset began with a plosive consonant (although we avoided bilabial to avoid visual cues preceding the plosion). In the stimuli recording session, all dialogues were written and displayed on a desk in front of the speakers during video shooting. Dialogues were all read by the same male and female speaker and videos were recorded in an anechoic room. To avoid body movement when reading the dialogues, the speakers sat on a chair and their head was maintained still by a hard hat fixed to the chair. Each speaker wore sunglasses to hide eye movements and a wig to hide the hard hat (Fig. 1A left). Furthermore, in order to discard spurious visual cues, the non-speaking speaker was "frozen" via video editing whenever a head or lip movement occurred before the turn. This happened in approximately ~20% of the turns. In order to avoid sharp jumps in the images of the upcoming speaker such freezing took advantage of fading and morphing techniques (mMorphCut plug-in; FINAL CUT pro X).

In other words, turns could not be anticipated on the basis of visual information of the next speaker right before the turn. Thus, anticipatory gaze behaviour cannot be interpreted in response to visual cues but as auditory predictive behaviour only. We also created 12 open questions for each dialogue (e.g. "why was Tom walking in the wood?") (see table 2 for more examples for one dialogue). These questions were recorded using an auditory only format, by a female speaker different from the one in the dialogues.

Speech recordings of the six dialogues were degraded using a noise vocoding approach (Shannon et al., 1995; custom Matlab script). This allows to parametrically vary the spectral detail, with increasing numbers of channels associated with increasing perceptual clarity (Fig. 1A right). The procedure allowed creating a highly degraded condition in which we kept only frequency bands from 120 Hz to 237 Hz, 405 Hz–538 Hz and 919 Hz–1028 Hz (i.e. 3 frequency bands left with 5 ERB scales) and a moderately degraded condition in which we removed all frequency bands except from 120 Hz to 237 Hz, from 674 Hz to 805 Hz, from 2025 Hz to 2112 Hz and 5208 Hz–5236 Hz (so 4 frequency bands left with 8 ERB scales). The overall RMS amplitude of the audio files was adjusted to be the same across all dialogues in the three conditions (normal, moderately degraded, highly degraded). We also run a pilot test to ensure that comprehension varied as a function of speech degradation and that, in the most challenging condition, participants could still understand part of the conversation. More precisely, 6 native French participants listened to audio excerpts of the dialogues and repeated all words they could recognize. The excerpts were degraded in a parametric manner in the number of channels, the bandwidth and the low frequency cut-off. The parameters were chosen in such a way to yield a condition of moderate difficulty (comprehension level between 30 and 50%) and another condition of great difficulty (comprehension level between 10 and 30%).

2.3. Procedure

Participants were equipped with a 64 preamplified Ag–AgCl electrodes (International 10/10 system site, BrainAmp system). The ground electrode was placed at AFz, the reference electrode at FCz, and the EEG signal sampling rate recording was 1000 Hz. Participants were comfortably seated in a Faraday sound-proof booth in front of a computer screen (24″) with a resolution of 1920 × 1080px and a refresh rate of 100 Hz), at a distance of approximately 70 cm. A Gazepoint GP3 eye-tracker (sampling rate: 60 Hz; accuracy: .5–1° of visual angle) was installed at the bottom of the screen on a tripod to record the participant's gaze positions. OpenSesame software (Mathôt et al., 2012) installed on a Dell laptop (Precision T1700) launched the eye-tracker recording through the PyGaze package (Dalmaijer et al., 2014) and the audio-visual stimulation. The videos were displayed on full screen resolution, and the sound was delivered through a 2040 YAMAHA amplifier and two NS 1020 Studio YAMAHA loudspeakers located on both sides of the screen. Before each video presentation, the participant's eyes were calibrated using the PyGaze's standard 12-points calibration procedure.

The experimenter asked participants to attentively follow the dialogues and informed them that, after each dialogue, they should answer 12 questions on the dialogue content. These questions allowed us to ascertain participants' attention and assess the level of comprehension. Each participant saw 2 dialogues in each of the three experimental conditions (normal speech, moderately degraded speech, highly degraded speech). The level of degradation of each dialogue and the order of the dialogues were counterbalanced across participants (each dialogue was presented at a different degradation level - normal, moderately degraded, highly degraded - to different participants).

2.4. Stimuli, eye-tracking and EEG data synchronization

A custom-made sync box based on an Arduino micro-controller ensured the synchronization between the EEG data and the acoustic stimulation that was embedded in the video file. The stereo sound goes from the stimulation PC to the sync box (audio cable). There, the stereo is split and the first channel containing the speech signals of the dialogues goes directly to the loudspeakers. The second channel signal goes through the micro-controller. This channel contains an

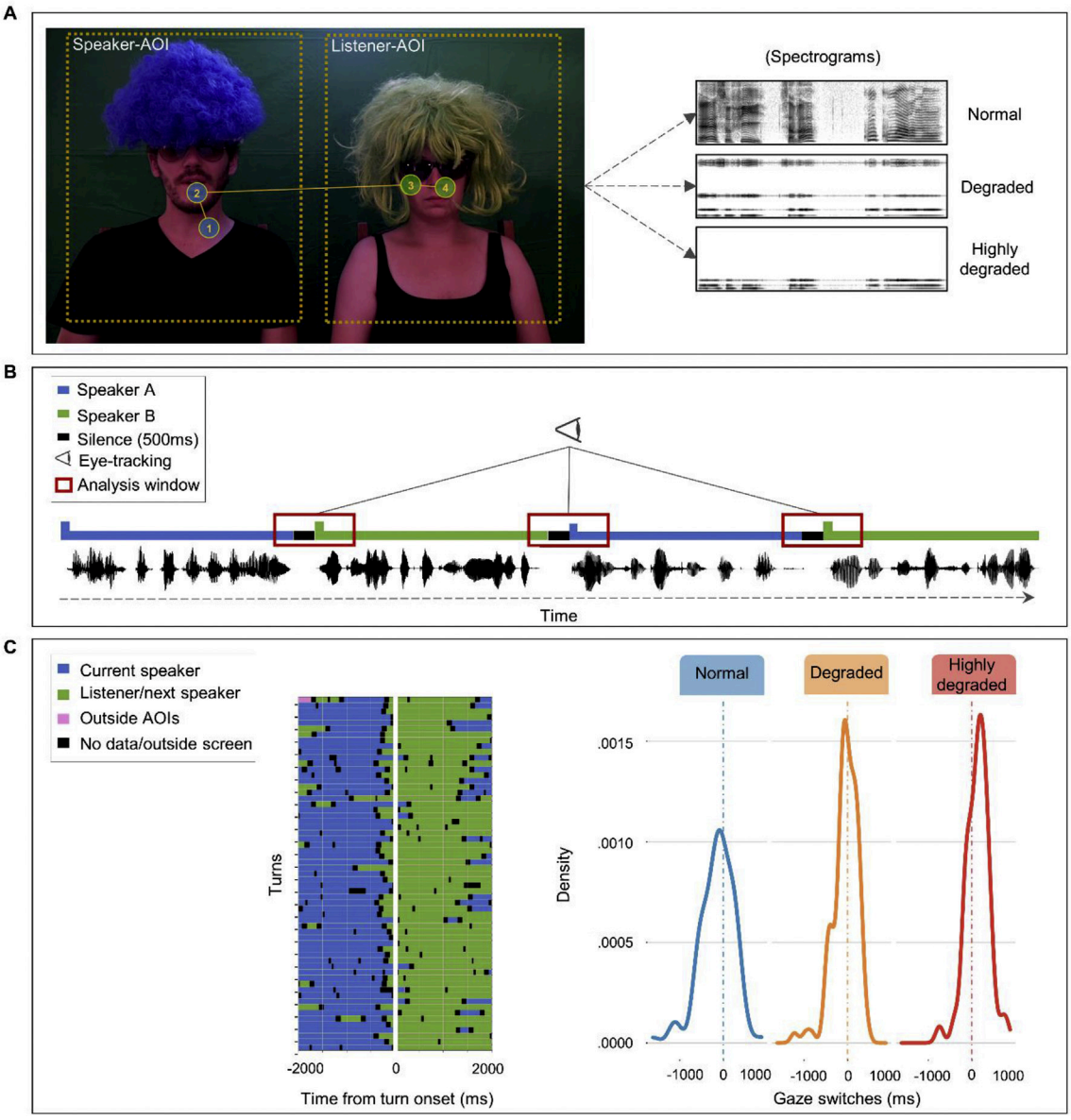


Fig. 1 – Schematic illustration of the task and analyses. A. On the left, snapshot of the video illustrating the fixation points of the participant and the gaze switch. On the right, the three different speech conditions with normal, moderately degraded and highly degraded speech (using a vocoding approach). B. Schematic illustration of the turns in the dialogue, the controlled gap between turns (500 ms) and the gaze switch window of analysis. C. On the left, gaze behaviour of a single participant showing the gaze switch for every turn of a single dialogue in the normal speech condition with respect to turn onset (zero). On the right, the gaze switch distribution for a single participant in the three different degradation conditions. A median split is used to class turns as being more or less well anticipated.

audio trigger indicating the beginning and the end of each dialogue. The Arduino detects the audio trigger and sends an adapted signal to the EEG system. To ensure a precise synchronization of the eye-tracking on the acoustic stimulation and the EEG data, we used saccadic movements. More precisely, saccades can be easily detected in the eye-tracking data, but they are also clearly visible in the EEG signal, especially using Independent Component Analysis decomposition (Makeig et al., 1996). Thus, for each dialogue, we detected gaze switch for every turn in both the eye-tracking and EEG data (see supplementary material 1). Then, we computed for each subject and dialogue the median delay between the two signals across all turns and used it to temporally realign the eye-tracking to EEG (median delay across participants = 23 ms; median sd across participants = 12 ms). This procedure ensured a good time-alignment of the three types of data (video, eye-tracking and EEG).

2.5. *Data analysis*

2.5.1. *Eye-tracking*

In this study we used the gaze switch from one speaker to the other as a proxy of anticipatory behaviour with respect to the upcoming turn onset (see Fig. 1A left and B). However, in

contrast with previous studies (Casillas & Frank, 2017) that were mostly interested in anticipatory gaze shift, we were also interested in somewhat late gaze shifts. We assumed that adult participants need approximately 200 ms to plan an eye movement (Kamide et al., 2003). Thus, we considered a temporal window around the turn allowing to keep both anticipatory and non-anticipatory gaze switch, resulting in a −1/+1s window around the turn onset (Fig. 1B). Note, nonetheless that, as expected, most shifts occurred in the inter-turn gap, that is before the turn onset (Foulsham et al., 2010; Keitel et al., 2013; Casillas & Frank, 2017, see Fig. 2B).

Moreover, we also applied three supplementary criteria to filter spurious gaze switches. First, gaze before and after the switch should fall within an area of interest (AOIs), defined as stationary rectangle surrounding each face (see Fig. 1A left). Second, gaze switch should be preceded by at least 100 ms fixation on the current speaker. Third, it should be followed by at least 150 ms fixation to the next speaker. Concerning the possible back and forth gaze behaviour preceding turn (4.5% in our data), we kept only the first gaze switch.

In short, we computed a gaze switch latency, that is relative to the upcoming turn onset time. A positive indicates thus a gaze switch following the turn onset, while a negative value indicates a gaze switch before the turn onset (Fig. 1C left). Then, we used a median split to class turns as early or late, separately for each subject and condition (Fig. 1C right).

2.5.2. EEG

Signal processing was done using EEGLAB (Delorme & Makeig, 2004) and custom scripts written in MATLAB. We high-pass filtered (0.5 Hz) continuous data and rejected major artifacts (>300 μV). For every participant, we systematically used Independent Component Analysis to remove eye blinks and saccadic movements and, when needed, muscular activity. We then low-pass filtered (45 Hz) and segmented continuous data into epochs of 900 ms starting at 100 ms prior to turn

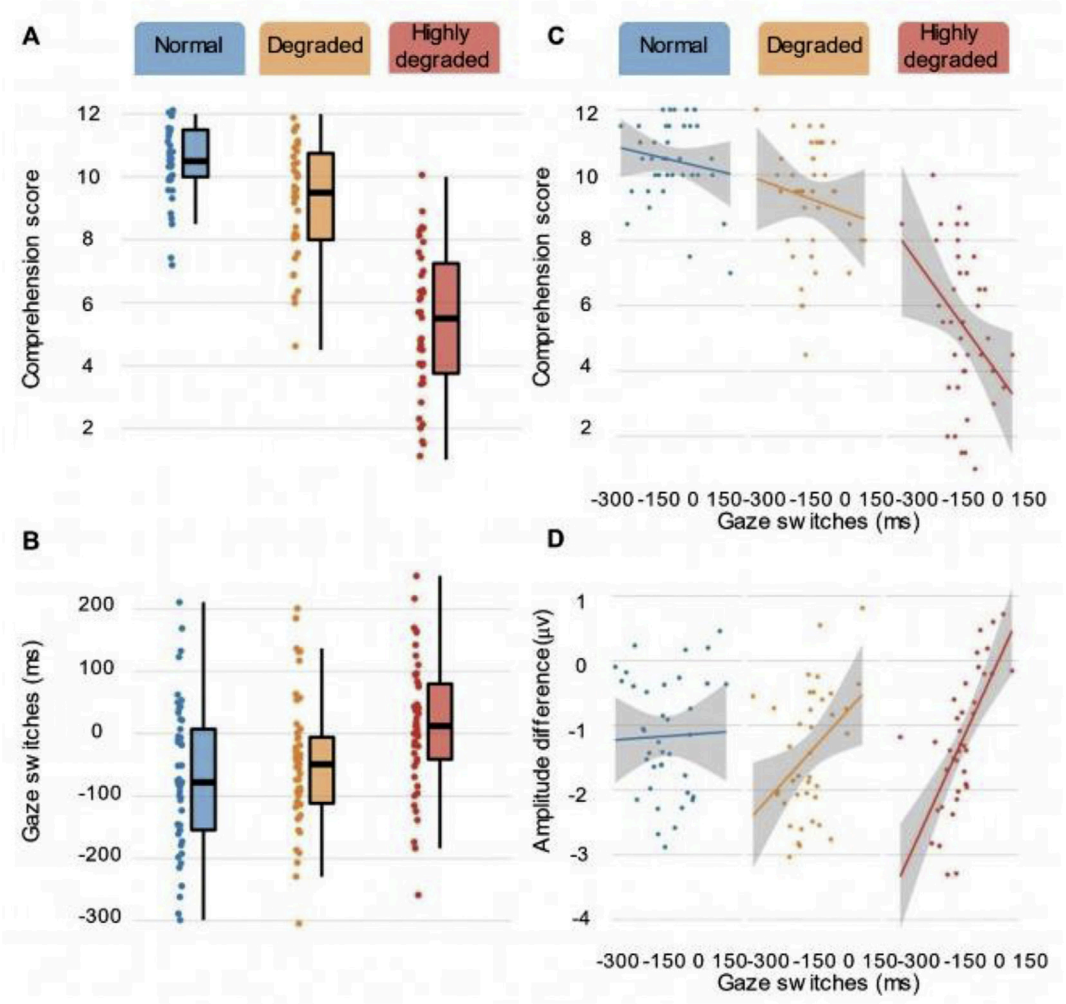


Fig. 2 – Behavioural results and correlations. A. Effect of speech degradation level on comprehension. Dots represent the average comprehension score for each participant in each condition. B. Effect of speech degradation level on gaze switch latency. Latency at zero milliseconds corresponds to the onset of the turn. Dots represent the average latency for each participant in each condition. C. Scatter plots and linear regression between comprehension scores and gaze switch latency in the three different conditions. D. Scatter plots and linear regression between N400 effect size (late minus early, in the 300–550 ms latency window) and gaze switch latency in the three different conditions.

onset. We zero-mean normalized epochs to the baseline ([−100, 0]ms) and re-referenced to the algebraic average of all electrodes. Finally, we averaged epochs according to the three conditions, also separating epochs according to the latency of the gaze shift corresponding to each turn (early or late, see above).

2.6. Statistical analysis

We computed all statistical analysis using R (Team, 2021) and the lme4 package (Bates et al., 2015). We evaluated participants' comprehension scores (1 point attributed to each good answer; range 0–12) as a function of speech degradation by fitting linear models on the mean of correct responses for each condition (2 dialogues by condition) and each subject. Similarly, we modelled the mean of gaze switch latencies for each dialogue as a function of speech degradation by fitting linear model on latency data. We then used a general linear model (glm) in modelling the time spent gazing at the current speaker (ratio of the total fixation time). All theses models were tested against the null model before being interpreted. We computed a linear mixed model with interaction to explain comprehension score according to gaze switch latencies and speech degradation, with subject as random effect (comprehension ~ gaze switch latency * speech degradation + 1|subject). This model was compared 1) to the null model, 2) to the model with only gaze switch latencies as predictor and 3) to the additive model (gaze switch latencies and speech degradation as independent predictors). Statistical significance of the fixed effects was assessed by model comparison using the Akaike Information Criterion, thus arbitrating between complexity and explanatory power of the models. Normality and homoscedasticity of the residuals of all the models were systematically visually inspected. Reported p values are Satterthwhaite approximations obtained with the lmerTest package (Kuznetsova et al., 2017).

The statistical significance of the differences between conditions for the ERP data was evaluated by a cluster-based random permutation approach for the full set of 64 electrodes (2-tailed test with 500 permutations over the whole time range of the ERP epoch, i.e. between −100 and 800 ms). This statistical approach handles the multiple-comparisons problem. More precisely, the approach controls for the Type-1 error rate in multiple testing across channels and time points by identifying clusters of significant differences between conditions in the time and space dimensions (Maris & Oostenveld, 2007). Finally, in order to assess the relation between gaze switch and ERPs, we used simple linear models with the amplitude difference between late and early trials in the N400 window explained as a function of the interaction between gaze switch latencies and speech degradation (N400 effect ~ gaze switch latency * speech degradation). As for the other analyses, this model was compared to the additive model.

3. Results

As expected, the comprehension score decreases as a function of increasing speech degradation ($\beta = -2.551$, SE = .22, $t = -11.55$, $p < .001$, Fig. 2A). Compared to the normal condition, both the highly degraded and degraded speech conditions altered the comprehension score ($\beta = -1.23$, SE = .42, $t = -2.929$ $p = .004$; $\beta = -5.102$, SE = .42, $t = -12.145$, $p < .001$, respectively).

It is important to note that participants' score are well above zero and this is also the case in the highly degraded condition (range: 1–10, median 5.5). Considering that the score is based on response to open questions, it is extremely unlikely that a single correct answer is given by chance. As such, these results show that participants of the current experiment can partly understand the dialogues even in the most difficult condition of speech perception.

When turning to eye-movements, participants gaze data are available on average for 87% of the total duration of the dialogues (range: [82–90]). The loss of data is possibly due to a gaze off the screen and/or to a temporary loss of the eye from the tracker. From these available data, it appears that participants gazed at the current speaker most of the time during video playback (93%; range: [87–97]). Compared to the normal condition, time spent on current speaker is greater in the degraded ($\beta = .580$, SE = .073, $t = 7.881$, $p < .001$) and highly degraded conditions ($\beta = .616$, SE = .074, $t = 8.317$, $p < .001$). Importantly, not only participants gazed most of the time at the current speaker, but when considering gaze switch towards the current speaker, these took place most often in a −1/+1sec window around the turn onset (mean 80%, range: [.56–.96]), the rest being gaze switches far away from the turn or missing data.

Normalizing by time unit this gives a number of gaze switch per second of .47 during the turn window and .08 during the rest of the dialogue (that is switches from one AOI to the other that were performed outside the −1/+1s window around the turn, see supplementary material 2). Overall, these results clearly show that gaze and gaze switches are not randomly distributed across the dialogues. On the contrary, gaze behaviour to current speakers demonstrates that participants are tracking turns during the videos, as previously reported (Casillas & Frank, 2017). When moving to gaze switch latencies, one can see that these are influenced by speech degradation ($\beta = 43.382$, SE = 9.52, $t = 4.553$, $p < .001$, see Fig. 2B). This is due to longer latencies to turns with highly degraded speech compared to normal and degraded speech (always $p < .001$) while there are no latency differences between degraded and normal conditions ($p > .1$). Moreover, latencies globally move from negative values in the normal condition to positive values in the highly degraded condition. Because latencies are estimated with respect to turn onset, this means that while in the normal condition participants realize the gaze switch before the turn, switches mostly take place after the turn onset in the highly degraded condition.

When looking at the relation between comprehension scores and gaze latency, this changes as a function of speech degradation ($\beta = -5.740$, SE = .001, $t = -3.936$, $p < .001$). More precisely, in the highly degraded condition shorter gaze switch latencies to turn onsets are associated to good comprehension scores ($\beta = -.008$, SE = .003, $t = -2.989$, $p < .01$). This effect was not significant in the moderately degraded nor in normal conditions ($p > .5$, Fig. 2C). This model with gaze switch latencies and speech degradation as interaction factors

was the best compared to the additive model (AIC = 1038.5 and 1043.6 respectively; p < .01 and p < .001).

The classification of turns on the basis of the latency of gaze switches (late vs early) shows a clear negative component for turns with late switches evolving between 350 and 700 ms post-stimulus onset over a fronto-central region (Fig. 3B). This effect is visible in the three different levels of speech degradation ($p < .05$, FDR corrected). However, this effect seems to differ across the three levels of speech degradation (Fig. 3A) insofar as the clusters start at different latencies. Because it is not appropriate to make an inferential claim on the time extension of the cluster (Sassenhagen & Draschkow, 2019), we run a further analysis using a jackknife approach in five subsequent 50 ms latency windows from 150 to 400 ms. This approach is considered appropriate to make inferences on the latency of an ERP effect (Miller et al., 1998). Results show that differences between early and late turns are significant starting 200 ms post turn onset in the normal condition, 150 ms in the moderately degraded condition and 300 ms in the highly degraded condition ($p < .05$, FDR corrected).

Finally, when looking at the relation between the ERP effect (late vs early gaze switches) in the 300–550 ms window and the global gaze switch latency of participants, we find a significant interaction between gaze switch latencies and speech degradation levels ($F = 7.135$, $p = .001$). More precisely, one can notice, that, in the highly degraded condition, participants with overall shorter gaze switch latencies to turn onsets show a larger effect ($\beta = .007$, SE = .001, $t = 3.753$, $p < .001$). A similar but not significant trend is also visible in the moderately degraded ($\beta = .003$, SE = .001, $t = 1.912$, $p = .058$) but not in the normal conditions ($\beta = .000$, SE = .001, $t = .204$, $p = .838$, Fig. 2D).

4. Discussion

In the present experiment, participants watched a series of videos showing two interlocutors having a conversation. We parametrically degraded speech and measured eye-movements, EEG and dialogue comprehension to gather a deeper understanding of the mechanisms underlying speech comprehension under challenging conditions. We build on the fact that participants switch gaze at the current speaker and use this measure as a proxy of prediction to class dialogue turns as more or less well predicted and study ERPs accordingly. Results show that 1) Participants switch gaze towards the current speaker in a limited temporal window around the turn. 2) Speech degradation reduces the anticipatory gaze behaviour. 3) Globally, participants with a low comprehension score tend to have a later gaze behaviour when speech is strongly degraded. 4) ERPs to turn onsets vary as a function of the latency of the gaze switch. 5) Finally, participants with overall shorter switch gaze latencies show a better comprehension score and a stronger anticipatory effect in their neural responses; this relation between behavioural and neural variables is significant only in the most challenging (degraded) condition. We discuss these findings with respect to the

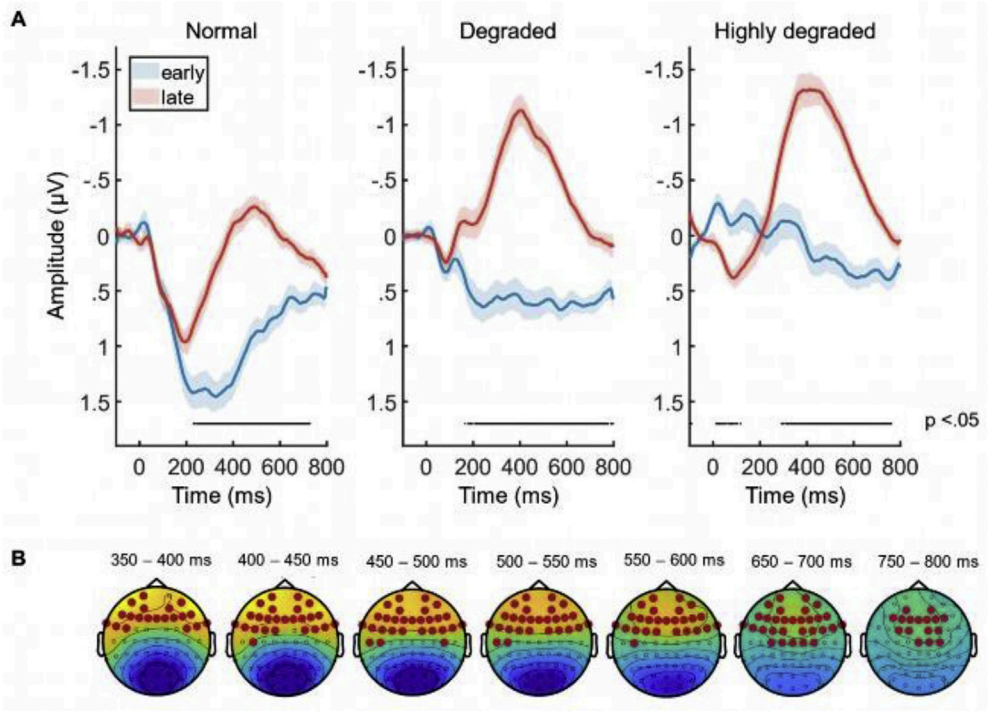


Fig. 3 – A. event-related potentials (ERPs) in a fronto-central region of interest time-locked to the turn onset for early and late gaze switch latencies. Shaded areas indicate the standard error of the mean. The black line below the ERPs indicates the FDR corrected significant difference between early and late conditions. B. Cluster-based statistical analyses on all electrodes showing the main effect of gaze behaviour (early vs late). The topography illustrates the difference between late and early turns across all levels of degradation. The red dots represent electrodes with significant differences (cluster corrected).

relation between predictions, neural predictions errors and speech comprehension under challenging conditions. Of course, these findings are limited by the constraints inherent to the present experiment, wherein predictions are possible purposely on the basis of auditory cues only because we controlled all extralinguistic visual cues (e.g. lip preparatory movement) that undoubtedly play an important role in conversation.

Our study replicates previous findings (Casillas & Frank, 2017; Foulsham et al., 2010; Keitel et al., 2013), showing that gaze switch takes place in a precise temporal window centered before turn onset. Considering the short period between turns and the time needed to prepare and realize a saccade, gaze switch, in this context, can be interpreted as anticipatory behaviour (Keitel & Daum, 2015). The new result, in this respect, is that gaze switch behaviour is affected by the more or less challenging speech comprehension, here obtained by spectrally degrading speech signal. The reduction of anticipatory gaze behaviour in presence of stronger speech degradation seems to indicate that increasing the cost of speech processing has a detrimental effect on the prediction of the upcoming turn. This is in line with results on hearing-impaired individuals showing a delay (greater effort) in processing sentences with more or less contextual cues, compared to normal-hearing (Winn, 2016).

Importantly, our results show a relation between comprehension score and turn anticipatory behaviour, under challenging conditions. This result is not trivial. Indeed, the comprehension score reflects the global understanding of the whole dialogues, because it corresponds to questions that were asked at the end of each dialogue. By contrast, gaze switch and EEG dynamics reflect anticipatory behaviour at precise temporal windows surrounding conversational turns. This means, that, under challenging conditions, participants that are able to understand can take advantage of an anticipatory gaze behaviour. Alternatively, participants with an anticipatory gaze behaviour are better able to understand. While the present design does not allow for a causal interpretation between gaze behaviour and global understanding of the dialogues, we will see that ERP results play here a very important role. However, before discussing this point one has to interpret the ERP components, in particular the negative component peaking around 400 ms.

First, one may raise the possibility that residual gaze-related artefact may still be present in the EEG, even following ICA procedures (see methods). However, 1) the systematic difference in latency between late and early switches was on average 480 ms, while we do not observe such a delay in ERPs to early and late turns; 2) while highly degraded speech induces overall later gaze switches, the N400 peak in this condition has the same latency as in the other conditions (in Fig. 3A the N400 peak is always 400 ms). Thus, overall, one can consider that differences in ERPs are due to cognitive processes and not to residual physiological artefacts due to eye movements.

The first and most likely interpretation of the larger negativity to both degraded turns and late turns is in terms of an N400 like component. Indeed, the amplitude of the N400 is strongly affected by expectations: the more a target word is unexpected within a sentence context, the larger the N400 amplitude (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984). While the N400 topography we report is more anterior compared to the classic centro-parietal distribution, this may be due to the audiovisual modality (Kelly et al., 2004) or to the presence of complex scenes (Kutas & Federmeier, 2011) inducing a topography more frontal than for written words. Importantly, the N400 has been proposed to be a proxy of prediction errors (Bornkessel-Schlesewsky & Schlesewsky, 2019; Rabovsky & McRae, 2014), that play a key role within the predictive coding framework (Friston, 2010; Schultz & Dickinson, 2000).

As stated above, in order to assess the relevance of predictive coding theories with respect to speech comprehension, it is necessary to have access to both prediction and prediction error indicators. Our results, as well as previous studies, seem quite clear in showing that gaze switch behaviour can be considered as a proxy of predictions. Previous research using an electrophysiological marker of prediction (the semantic prediction potential, SPP) showed that smaller SPP amplitude (weak predictions) result in a large N400 (Grisoni et al., 2021). Similarly, our results show that turns preceded by a late gaze switch, indicating poor predictions, give rise to a larger N400 like complex compared to turns preceded by an early gaze switch. An interesting advantage of the current design is that the target words (here turn onsets) do not need to differ in content or in context in order to engender a N400. This is different for instance, from previous studies manipulating the semantic and temporal turn relation to induce an N400 effect (Bögels et al., 2015). In other words, in the current study, the word and the context eliciting or not the N400 are strictly identical and we take advantage of the natural variability in the accuracy of participants' predictions across turns.

The interpretation in terms of N400 also fits quite well with the literature on weaker expectancies under degraded speech. Previous work showed that adverse listening conditions narrow the expectancies about the upcoming speech. This is visible in a reduced N400 effect in response to incongruent or less likely words under acoustic degradation (Aydelott et al., 2006; Obleser & Kotz, 2011; Strauß et al., 2013). The present findings show a somewhat similar result, with an N400 effect that is significant in a later latency window in the most challenging condition (strong speech degradation) compared to the two other conditions. While the typical N400 effect concerns the amplitude of the negative wave, several works reported later N400 latencies in context presenting more difficult semantic access (Deacon et al., 1995; Moreno & Kutas, 2005). Interestingly, the temporal window analysis shows that traces to early and late turns diverge first in the degraded condition, then in the normal and last in the highly degraded conditions. It thus seems as if the advantage of an early gaze switch was integrated faster in moderately challenging conditions. There, the benefit of an anticipatory gaze behaviour and temporally precise (local) predictions may be maximal compared to the two other conditions wherein uncertainty may either be too low (normal speech) or too high (highly degraded speech).

However, while the N400 latency is earliest in the degraded condition (Fig. 3A), the N400 amplitude best correlates with turn anticipatory behaviour in the highly degraded condition (Fig. 2D). In considering this inconsistency one should keep in

mind that 1) the N400 effect, in terms of amplitude, is also significant in the highly degraded condition; 2) while the correlation between N400 amplitude and average gaze behaviour does not reach the significance threshold ($p = .06$), the trend of the correlation is similar across the two degraded conditions.

Importantly, Fig. 2C and D clearly show that there is a similar relation between gaze behaviour and the semantic access of dialogues. Differently from global comprehension score, the N400 amplitude does not reflect a global understanding of the conversation, but it is a rather local measure, as it is the case with gaze switch behaviour. Moreover, the two measures are serially ordered in time, with gaze switch taking place earlier ~400 ms before the N400 peak. Thus, one can hypothesize that participants with an overall early gaze switch, by having a better prediction of the turn, have an easier lexical-semantic integration of the word starting the turn in the conversational context. This relation between gaze and both local (N400 effect) and global (comprehension scores) understanding is only shown under challenging conditions, wherein indeed participants can benefit of a predictive behaviour to improve their understanding of the dialogues.

It seems appropriate here to evoke an alternative (although not necessarily the most likely) interpretation of the negative component in terms of a phonological mismatch negativity (pMMN). Several studies have addressed the timing of audio-visual integration. For instance, in McGurk illusion, the temporal window allowing modality fusion and allusion ranges between 30 ms and 170 ms of asynchrony (van Wassenhove et al., 2007). When looking at the electrophysiological response to the incongruent audio-visual stimulation, several authors report a pMMN (Colin et al., 2002; Eskelund et al., 2015; Stekelenburg & Vroomen, 2012). The latency of this response is later than the classic MMN and is described approximately 500 ms after the voice onset (Proverbio et al., 2018). In our study, the classification of turns as being accompanied by late or early gaze switch implies that the audio-visual relation of the stimuli differs in the two classes. For early gaze switch, the gaze is well anticipated and thus the audio-visual information is perceived simultaneously. By contrast, in turns with a late gaze switch, often occurring after the voice onset, the audio-visual integration has a higher level of uncertainty. In other words, if a gaze switch takes place later than a turn onset, there may be an audiovisual mismatch because the participant is looking at the previous speaker while listening to the upcoming speaker's voice. In this context, looking at the closed lips of the previous speaker and hearing the word pronounced by the following speaker, may engender a phonological audiovisual MMN.

We would like to end the discussion with a consideration on the implicit and explicit role of predictions. It is known that the influences of the visual system on auditory perception can be so strong as to override under certain conditions the original input of the auditory system (Mcgurk & Macdonald, 1976). More commonly, viewing the lips provides relevant complementary information and can augment and improve auditory capabilities (Calvert et al., 1997; Drijvers & Özyürek, 2017; Grant & Seitz, 2000; Sumby & Pollack, 1954). In this context, it is interesting to note that, compared to the normal condition, the moderately degraded condition yields a poorer comprehension but similar gaze switch latencies and an earlier N400 effect. These preserved latencies, in a context of poorer comprehension that should rather induce poorer predictions, may be mediated by an active compensatory strategy, explicitly making greater use of visual cues to improve predictive processes (Sohoglu & Davis, 2016). Such a compensatory strategy may not hold when speech is highly degraded, yielding to slower gaze switches due to a poor understanding of the semantic and syntactic context. The respective role of implicit and explicit predictions and their link to conversation strategies (Hadley et al., 2021) will require further work addressing this specific question. Another interesting aspect that will require further work is the inter-relation between local predictions (gaze), local semantic processing (N400) and global comprehension (here assessed via specific questions on the dialogues). While on one side gaze switches temporally precede auditory and semantic processing of the upcoming turn, this anticipatory behaviour can only build on the global comprehension of the dialogue, that, in turn, depends on the integration of local semantic processing.

To conclude, this study confirms that gaze switch can be used as a proxy of predictions in a conversational context. It shows that these predictions are related to the lexico-semantic processing of the turn start, as estimated by a neural marker of prediction error. It also shows that predictions are less accurate in more challenging listening conditions, but that they are also most useful in that specific context to make sense of the upcoming turn and, more generally, are a good indicator of global dialogue comprehension.

Author note

In accordance with the Peer Reviewers' Openness Initiative (https://opennessinitiative.org, Morey, Chambers, Etchells, Harris, Hoekstra, Lakens, et al., 2016), all materials and scripts associated with this manuscript are available on https://osf.io/7dc6y/

Funding sources

Work supported by APA foundation (RD-2016-9), ANR-16-CONV-0002 (ILCB) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

Author contributions

Conceptualization C.H. and D.S.; Data curation C.H., C.Z. and I.M.; Formal Analysis C.H.; Funding acquisition D.S.; Project administration C.H.; Supervision D.S.; Visualization C.H, C.Z. and I.M; Writing – original draft C.H., I.M., C.Z. and D.S.

Open Practice

The study in this article earned Open Data and Open Materials badges for transparent practices. Stimuli will be available on

individual on simple request to the senior author Daniele Schön. Readers seek access to the stimuli are advised to contact the senior author Daniele Schön. Stimuli will provided on request without restriction by writing an email to daniele.schon@univ-amu.fr.

Acknowledgements

We would like to thank Laura Leone, Romane Pradels and Benjamin Morillon for their participation in the design of the dialogues and the production of videos.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2022.01.020.

REFERENCES

Aydelott, J., Dick, F., & Mills, D. L. (2006). Effects of acoustic distortion and semantic context on event-related potentials to spoken words. *Psychophysiology, 43*(5), 454–464. https://doi.org/10.1111/j.1469-8986.2006.00448.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using {lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beňuš, Š., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics, 43*(12), 3001–3027. https://doi.org/10.1016/j.pragma.2011.05.011

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no . . . How the brain interprets the pregnant pause in conversation. *Plos One, 10*(12). https://doi.org/10.1371/journal.pone.0145474

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology, 10*(FEB), 298. https://doi.org/10.3389/fpsyg.2019.00298

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593–596. https://doi.org/10.1126/science.276.5312.593

Casillas, M., & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language, 92*, 234–253. https://doi.org/10.1016/j.jml.2016.06.013

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology, 113*(4), 495–506. https://doi.org/10.1016/S1388-2457(02)00024-X

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods, 46*(4), 913–921. https://doi.org/10.3758/s13428-013-0422-2

De Ruiter, J., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82*, 515–535. https://doi.org/10.1353/lan.2006.0130

Deacon, D., Mehta, A., Tinsley, C., & Nousak, J. M. (1995). Variation in the latencies and amplitudes of N400 and NA as a function of semantic priming. *Psychophysiology, 32*(6), 560–570. http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=11064097&lang=fr&site=ehost-live.

Delorme, A., & Makeig, S. (2004). Eeglab: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods, 13*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research, 60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Eskelund, K., MacDonald, E. N., & Andersen, T. S. (2015). Face configuration affects speech perception: Evidence from a McGurk mismatch negativity study. *Neuropsychologia, 66*, 48–54. https://doi.org/10.1016/j.neuropsychologia.2014.10.021

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition, 117*(3), 319–331. https://doi.org/10.1016/j.cognition.2010.09.003

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197. https://doi.org/10.1121/1.1288668

Grisoni, L., Tomasello, R., & Pulvermüller, F. (2021). Correlated brain indexes of semantic prediction and prediction error: Brain localization and category specificity. *Cerebral Cortex, 31*(3), 1553–1568. https://doi.org/10.1093/cercor/bhaa308

Hadley, L. V., Whitmer, W. M., Brimijoin, W. O., & Naylor, G. (2021). Conversation in small groups: Speaking and listening strategies depend on the complexities of the environment and group. *Psychonomic Bulletin & Review, 28*(2), 632–640.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Keitel, A., & Daum, M. M. (2015). The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in Psychology, 6*(February), 108. https://doi.org/10.3389/fpsyg.2015.00108

Keitel, A., Prinz, W., Friederici, A. D., Hofsten, C. von, & Daum, M. M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *Journal of Experimental Child Psychology, 116*(2), 264–277. https://doi.org/10.1016/j.jecp.2013.06.005

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161–163.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Levinson, S. C. (2016). Turn-taking in human communication - origins and implications for language processing. *Trends in Cognitive Sciences, 20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 145–151.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748. https://doi.org/10.1038/264746a0

Miller, J., Patterson, T. U. I., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology, 35*(1), 99–115.

Molinaro, N., Monsalve, I. F., & Lizarazu, M. (2016). Is there a common oscillatory brain mechanism for producing and predicting language? *Language, Cognition and Neuroscience, 31*(1), 145–158. https://doi.org/10.1080/23273798.2015.1077978

Moreno, E. M., & Kutas, M. (2005). Processing semantic anomalies in two languages: An electrophysiological exploration in both languages of Spanish-English bilinguals. *Cognitive Brain Research, 22*(2), 205–220. https://doi.org/10.1016/j.cogbrainres.2004.08.010

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage, 55*(2), 713–723. https://doi.org/10.1016/j.neuroimage.2010.12.020

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing, 39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Proverbio, A. M., Raso, G., & Zani, A. (2018). Electrophysiological indexes of incongruent audiovisual phonemic processing: Unraveling the McGurk effect. *Neuroscience, 385*, 215–226. https://doi.org/10.1016/j.neuroscience.2018.06.021

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition, 132*(1), 68–89. https://doi.org/10.1016/j.cognition.2014.03.010

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*(3), 618–660. https://doi.org/10.1037/0033-2909.85.3.618

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology, 6*, 509. https://doi.org/10.1093/cercor/bhl024. https://doi.org/10.3389/fpsyg.2015.00509.

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology, 56*(6), e13335.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23*, 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA), 6*(3), 371–388. https://doi.org/10.1075/prag.6.3.06sel

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. In *Proceedings of the national academy of sciences of the United States of America* (Vol. 113, pp. E1747–E1756). https://doi.org/10.1073/pnas.1523266113 (12).

Šoškić, A., Jovanović, V., Styles, S. J., Kappenman, E. S., & Ković, V. (2021). How to do better N400 studies: Reproducibility, consistency and adherence to research standards in the existing literature. *Neuropsychology Review*, 1–24.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass, 9*(8), 311–327. https://doi.org/10.1111/lnc3.12151

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience, 6*(MAY 2012), 1–7. https://doi.org/10.3389/fnint.2012.00026

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. In *Proceedings of the national academy of sciences of the United States of America* (Vol. 106, pp. 10587–10592). https://doi.org/10.1073/pnas.0903616106 (26).

Strauß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed expectancies under degraded speech: Revisiting the N400. *Journal of Cognitive Neuroscience, 25*(8), 1383–1395. https://doi.org/10.1162/jocn_a_00389

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212–215. https://doi.org/10.1121/1.1907309

Team, R. C. (2021). *R: A language and environment for statistical computing*. Vienna: R Project, 2017.

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34*(5), 409–420.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

Winn, M. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing, 20*, 1–17. https://doi.org/10.1177/2331216516669723