Aix*Marseille université
Socialement engagée

NNT/NL : 0000AIXM0000/000ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 8 décembre 2023 par

Charly LAMOTHE

Du signal vocal à la réponse neuronale : modélisation computationnelle et perspective évolutive sur le traitement de la voix dans le cortex auditif

Discipline
Biologie – Santé

Spécialité
Neurosciences

École doctorale
ED 62 – Sciences de la Vie et de la Santé

Laboratoire/Partenaires de recherche
Institut des Neurosciences de la Timone – UMR 7289
Laboratoire d'Informatique et Systèmes – UMR 7020

Composition du jury

Kate WATKINS	Rapporteure
Professeure, Université d'Oxford, Oxford	
Jean-Julien AUCOUTURIER	Rapporteur
Directeur de recherche, CNRS UMR 6174, Besançon	
Ian CHAREST	Examinateur
Professeur adjoint, Université de Montréal, Montréal	
Ricard MARXER	Examinateur
Professeur, Université de Toulon, Toulon	
Timothée PROIX	Examinateur
Docteur, Université de Genève, Genève	
Pascal BELIN	Directeur de thèse/Président du jury
Professeur, AMU, Marseille	
Thierry ARTIÈRES	Co-directeur de thèse
Professeur, Ecole Centrale Marseille & AMU, Marseille	
Etienne THORET	Membre invité
Chargé de recherche, CNRS UMR 7289 & AMU, Marseille	
Brice BATHELLIER	Membre invité
Directeur de recherche, CNRS, Paris	

int institut de neurosciences de la timone
LIS LABORATOIRE D'INFORMATIQUE & SYSTÈMES
erc European Research Council Established by the European Commission
ILCB

Affidavit

I, undersigned, Charly Lamothe, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Pr. Pascal Belin and Pr. Thierry Artières 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.

Marseille, September 6th 2023

List of publications and participation in conferences

List of publications conducted within the framework of the thesis project:

1. **Lamothe C.***, Cordeau M.*, Obliger-Debouche M.*, Giamundo M., Artières T. Belin P. (*to be submitted*). The primate cerebral "voice patches": organization, evolution, computation.
2. **Lamothe C.*** & Obliger-Debouche M.*, Best P., Trapeau R., Ravel S., Artières T., Marxer R., Belin P. (*to be submitted*). MarmAudio: A large annotated dataset of the common marmoset vocal repertoire vocal repertoire.
3. **Lamothe C.**, Thoret E., Trapeau R., Giordano BL., Sein J., Takerkart S., Ayache S., Artières T., Belin P. (*submitted*). Reconstructing voice identity from auditory cortex using deep learning.

Participation in conferences and summer schools during the thesis period:

1. *3rd ILCB Summer School*, Marseille, 2020.
2. *26th Annual Meeting of the Organization of Human Brain Mapping*, Online, 2020.
3. Poster presentation at *NeuroFrance 2021*, Online, 2021.
4. Poster presentation at *27th Annual Meeting of the Organization of Human Brain Mapping*, Online, 2021.
5. *4th ILCB Summer School*, Marseille, 2021.
6. Invited talk at *School of Information Technology*, Halmstad University, Online, 2021.
7. Invited talk at *'Human, Monkey, Cetacean Communication Signals What to compare? How?' Workshop*, IMéRA, Aix-Marseille University, Marseille, France, 2022.
8. Poster presentation at *FENS Summer School on 'Artificial and natural computations for sensory perception: what is the link?'*, Bertinoro, Italy, 2022.
9. Poster presentation at *28th Annual Meeting of the Organization of Human Brain Mapping*, Glasgow, Scotland, 2022.
10. Poster presentation at *2nd European Workshop on Marmoset Neuroscience*, Marseille, France, 2022.
11. Poster presentation at *FENS Forum 2022*, Paris, France, 2022.
12. Invited talk at *European Center for Research in Medical Imaging*, Aix-Marseille Université, Marseille, 2023.
13. Poster presentation at *NeuroFrance 2023*, Lyon, France, 2023.
14. Poster presentation at *29th Annual Meeting of the Organization of Human Brain Mapping*, Montréal, Canada, 2023.
15. Invited talk at *CONECT Workshop "Active Learning in Brain and Machine"*, La Timone Neuroscience Institute, Aix-Marseille Université, Marseille, 2023.

Résumé

Les voix sont omniprésentes dans notre quotidien, essentielles pour communiquer et véhiculant des informations non verbales telles que l'identité, le genre ou l'émotion de nos congénères. De nombreuses espèces manifestent des aptitudes vocales sophistiquées. L'étude des mammifères, notamment des primates non humains, révèle des traits communs de traitement vocal. L'objectif de cette thèse est d'approfondir notre compréhension des mécanismes de traitement de la voix, en employant des méthodes computationnelles, y compris l'apprentissage profond, pour mieux comprendre l'encodage vocal et la communication vocale. L'introduction offre une vue d'ensemble des primates, leur évolution et communication, et introduit les concepts clés et méthodologies pour les chapitres suivants. Le premier chapitre présente une analyse comparative approfondie du cortex vocal chez les primates, et défini les limites de notre connaissances du traitement de l'information vocal par le cerveau. Le deuxième chapitre traite de l'utilisation de méthodes computationnelles pour construire une base de données de vocalisations de primates non humains, offrant une ressource précieuse pour les futures études en neuroéthologie. Enfin, le troisième chapitre examine la corrélation entre l'activité cérébrale liée à l'identité vocale, mesurée à l'aide de techniques de neuroimagerie, et les représentations issues de l'apprentissage profond via l'encodage et le décodage. Associant neuroimagerie, modélisation computationnelle et base de vocalisations, cette thèse enrichit notre compréhension du traitement vocal des primates, éclairant les origines du langage humain et offrant de nouvelles perspectives en neurosciences auditives.

Mots-clés : aires vocales, perception vocale, apprentissage profond, IRMf, base de vocalisations.

Abstract

Voices are ubiquitous in our daily surroundings, essential for communication, and rich in non-verbal information, such as the identity, gender, or emotional state of our conspecifics. Various forms of vocal communication are evident across species, with many demonstrating sophisticated vocalization capabilities. An examination of mammals, particularly non-human primates, indicates shared voice processing traits. This thesis aims to deepen our understanding of voice processing mechanisms, employing computational methods, including deep learning, to shed light on voice encoding and vocal communication. The introduction provides an overview of primates, their evolution, and communication methods, as well as introduces the fundamental concepts and methodologies to provide the necessary knowledge for understanding the subsequent chapters. The first chapter delivers an in-depth comparative analysis of the vocal cortex in primates and draws the limits of our current knowledge on voice processing in the brain. The second chapter addresses the employment of computational methods to build a dataset of non-human primate vocalizations, providing a valuable resource for future studies in neuroethology. Lastly, the third chapter examines the correlation between brain activity related to voice identity, as measured using neuroimaging techniques, and representations derived from deep learning through encoding and decoding. This thesis significantly augments our grasp of primate auditory vocal processing by combining neuroimaging tools, computational modeling, and a comprehensive vocalization database. The insights gained offer a deeper understanding of the evolutionary precursors of human vocal communication and present new opportunities for auditory neuroscience research.

Keywords: voice areas, voice perception, deep learning, fMRI, vocalization database.

Acknowledgements

Je souhaite exprimer ma profonde reconnaissance envers mon directeur de thèse, Pr Pascal Belin, ainsi qu'à mon co-directeur, Pr Thierry Artières, pour leur encadrement, soutien et conseils judicieux tout au long de ma thèse. Leur expertise complémentaire a été cruciale pour l'accomplissement de ce travail, et j'ai été honoré de bénéficier de l'opportunité de travailler sous leur direction. Je leur suis reconnaissant pour les connaissances et compétences qu'ils ont partagé avec moi, ainsi que pour leur mentorat bienveillant et attentif à mes besoins. Leur confiance m'a été précieuse. Merci à Pascal pour la liberté qu'il m'a accordée durant ma thèse, m'offrant ainsi l'occasion de m'épanouir et d'acquérir de l'indépendance. Merci à Thierry pour la rigueur qu'il m'a engagé à avoir dans mes nombreuses expériences.

I would like to express my sincere gratitude to the members of my thesis committee: Pr Kate Watkins, Dr. Jean-Julien Aucouturier, Dr. Ian Charest, Pr Ricard Marxer, Dr. Timothée Proix, Dr. Etienne Thoret, Dr. Brice Bathellier.

Je tiens également à remercier mes collaborateurs : Dr Etienne Thoret, Dr Bruno Giordano, Dr Manon Obliger-Debouche, Dr Margherita Giamundo, Dr Mélina Cordeau, Dr Régis Trapeau, Dr Paul Best, Pr Ricard Marxer. Un merci particulier à Bruno et Etienne pour leur mentorat et leurs conseils précieux tout au long de ma thèse, qui m'ont véritablement aidé à me développer en tant que chercheur. Merci de m'avoir traité comme un chercheur à part entière et de m'avoir fait confiance dès le début de ma thèse.

Merci à mes parents et ma sœur pour leur amour et soutien inconditionnels, et pour me rappeler souvent qu'ils m'aiment et sont fiers de moi. L'écriture de ma thèse n'aurait jamais été possible aussi rapidement sans eux.

Merci à ma famille au Québec pour leur accueil chaleureux, comme si nous nous étions quittés la veille, et pour tous ces moments précieux passés ensemble : Jean-Luc, Isabelle, Louis, Sylvie, Guy, Emile, Mariah, Julie, Simon, Jacynthe, Mélodie, Fannie, Michel, Jean-Denis, Stéphanie, et tous les petits cousins.

Merci à mes colocs qui sont devenus une deuxième famille : Alexus, Mélinight, Laura, Margux, Marie bras.

Merci à tous mes amis et collègues à l'INT et ILCB d'avoir rendu ces années de doctorat beaucoup plus amusantes que prévues : Mélina, Manon, Alex, Steven, Hannah, Damiano, Lina, Maxime, Davidou, Yoan, Marguerite, Marie sanguine, Marie bras, Etienne, Tiphaine, Ruggero, Etienne, Tiphaine, Maud, Sneza, et tous les autres que j'oublie.

Merci à l'équipe BaNCo pour tous ces moments conviviaux chez Pascal et aux labs meetings : Pascal, Bruno, Etienne, Thierry, Manon, Sylvain, Mélina, Régis, Margherita, Lina, Clem, Fati, Qi.

Merci aux membres passés et presents de l'équipe Qarma du LIS, et notamment merci à Thierry, Stéphane, Céline, FX, Valentin, Hachem, Ronan, Léo, Balthazar, Farah, Mimoun, Swetali, Hamed, Raphael, Hanwei, Rohit, Sokol, Marina, Julien, Luc, Riikka, Qi, Akrem, Paolo.

Merci à l'Indian cooking team pour ces moments conviviaux, à la coloc et à Malmousque, Julia bleue, Swetali, Léo, Rohit, Manvi, Khushboo, Shrabasti. Merci aux amis Despicabless, Swetali, Léo, Hamed, Julia, Hanwei, Mimoun. Merci aux amis qui sont loin de Marseille mais proches de mon cœur, Ihsane, Xiao Rui, Hanwei, Kep-Kee, Sophie, Haochen, Tiago, Misha, Evie, Gaetan, Maryline, Sara, Zhiyi. Merci pour votre soutien, surtout dans mes moments les plus bas. Merci aux amis Bikers du dimanche, Raphael, Julien, Vi, Katya. Merci à Dr Dona Awa et Don Leonardo de m'avoir accueilli dans votre gang de joyeux fêtards. Merci à Pauline la cous. Thank you Céline for the PhD meme support during the thesis writing.

Merci aux participants de mes études de m'avoir prêté leurs magnifiques cerveaux : Eva, Mélinda, Magali.

Merci aux colocs du Roi René : Antoine, Salomé, Marine, Oscar, Alex.

Merci aux grimpeurs.

Merci à ceux que j'ai oublié.

Merci à tous, je vous aime <3.

Preface

Voices permeate our daily lives, presenting in diverse forms such as speech, song, laughter, and emotional expressions. Voices are the carrier of speech, yet they convey extensive non-verbal information, hinting at aspects like the speaker's species, age, gender, emotions, and personality traits. Such vocal channels, transcending mere human speech, are common across numerous species. Many species excel in generating intricate vocalizations and decoding the information they convey. Given the emphasis on the neural underpinnings of vocal communication in recent theories, understanding voice processing's computational and evolutionary aspects in the auditory cortex is crucial.

The main goal of my thesis is to deepen our understanding of voice processing mechanisms. It emphasizes studying the vocal cortex in primates, touching on its functional and anatomical dimensions, and leveraging computational techniques, notably deep learning, to shed new light on voice encoding and vocal communication mechanisms.

This thesis is organized into four chapters, each investigating a specific facet of voice cerebral processing in primates. The introduction overviews primates, their evolutionary trajectory, and communication modalities. It also introduces the fundamental concepts and methodologies to provide the necessary knowledge for understanding the subsequent chapters. The second chapter delivers an in-depth comparative analysis of the vocal cortex in primates. The third chapter employs computational methods to build a large dataset of non-human primates' vocalizations. In the third chapter, I examine the correlation between brain activity related to voice identity—captured through neuroimaging—and the representations derived from deep learning. Finally, I synthesize established literature with the novel findings from my research in a general discussion.

Contents

Introduction

1. Motivation

Voices are a constant feature in our everyday environments. We encounter the voices of others in various forms, whether speaking, singing, laughing, or expressing emotions in other vocal ways. For most people, the primary function of the voice is to facilitate speech and language, one of the most advanced forms of interpersonal communication. Beyond this, voices convey extensive non-verbal information. They can hint at the speaker's species and identity attributes such as gender and age, emotional states like joy or sorrow, and even personality nuances. Unlike unique human speech, many of these channels of vocal communication are shared across various species. Numerous species have refined their ability to produce complex vocalizations and have developed the cognitive and neural capabilities to interpret the information in these vocal signals. Non-human primates, our closest evolutionary relatives, show comparable patterns in processing vocal information, both at the behavioral and neurological levels. A comparative approach, in which data from humans and non-human primates inform each other, is particularly promising, as it yields valuable information about the evolution of communication systems.

Although the anatomical-functional pathway supporting sound processing across primate species is well understood (see *Evolution of primates*), our knowledge of how the brain transforms species-specific vocal signals into meaningful semantic representations needs to be better defined. In particular, ***how is voice identity encoded in the brain?*** This primary question has guided my thesis, aiming to deepen our understanding of the intricacies of voice processing mechanisms. In Chapter 1, I first draw upon prior research on comparative voice perception to pose the question, ***what is the functional role of each unit within the "voice patch" system in the primate brain when processing vocal information?*** Reviewing older and recent literature on voice processing in humans and non-human primates (macaque, marmoset) (Belin et al., 2018; Bodin & Belin, 2020), I propose a synthesized functional model for voice information processing.

Nevertheless, it is still being determined how they process identity information and **whether or not voice patches across the primate species share similar coding principles**. I address this challenge within the neuroimaging paradigm, which consists of scanning humans and non-human primates while listening to conspecific vocalizations to model the relationship between voice signal properties and vocal brain activity using computational models (Figure 1). Merging our insights on conspecific voice processing with these computational models offers a new path to understanding how the brain represents and transforms the voice. With the rise of theories emphasizing the neural foundation for vocal communication, a deep exploration of the computational and evolutionary perspectives of voice processing in the auditory cortex becomes essential to understand better how the brain processes vocal communication. Building on research that models neural representations of speech or language with deep neural networks (DNNs) (Kell et al., 2018; Millet et al., 2022; Caucheteux et al., 2022; Caucheteux & King, 2022; Caucheteux et al., 2023; Giordano et al., 2023), I ask, ***would DNNs provide reasonable approximations of cerebral representations?*** — in particular regarding the processing of **voice identity**. To investigate this question across primate species, one would need to feed DNNs with a sufficient number of conspecific vocalizations (100,000s of samples) — a class of models that excel at learning high-level representations proportionally to the dataset size (LeCun et al., 2015) — as well as a sufficient number of corresponding vocal brain responses (10,000s of samples) to reliably span the voice space and reduce the notoriously high signal-to-noise ratio in neuroimaging techniques (e.g. in fMRI: Welvaert & Rosseel, 2013).

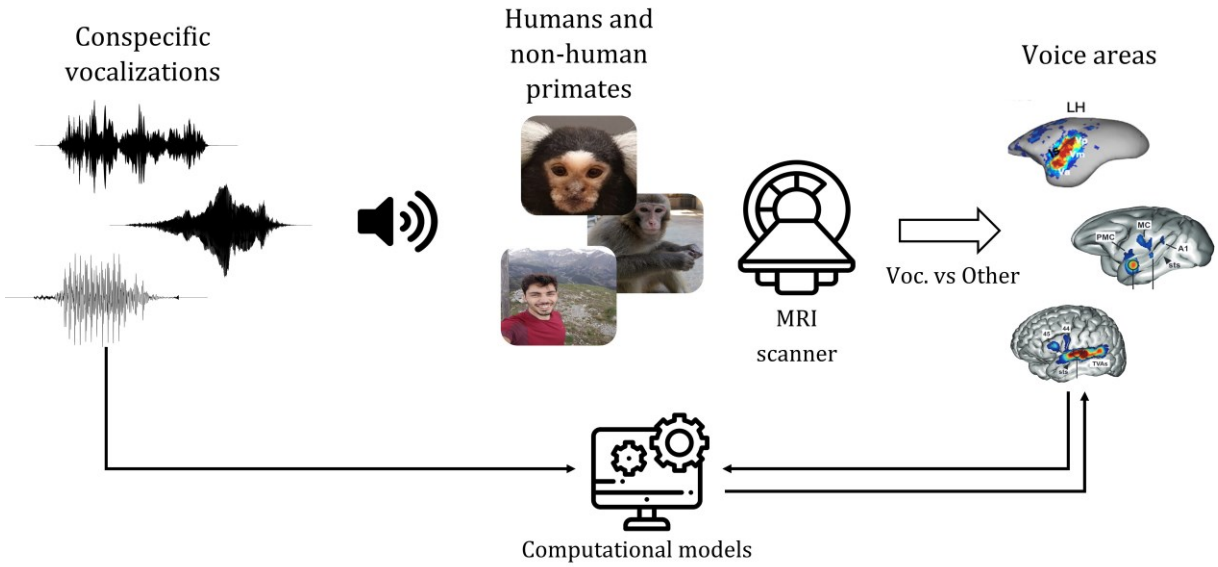


Figure 1: General paradigm used in this manuscript. The neuroimaging paradigm consists of scanning humans and non-human primates while listening to conspecific vocalizations to model the relationship between voice properties and voice areas. For example, the brain activity is recorded using functional magnetic resonance imaging (fMRI), and the voice areas are estimated by computing the contrast 'Voc. vs. Other', i.e., conspecific vocalizations brain activity vs. other vocalizations and sounds brain activity.

My Ph.D. project focuses on two aspects of this paradigm. There is yet to be a large and labeled dataset of monkey vocalizations. To fill this gap, I describe in Chapter 2 an end-to-end pipeline for processing vocalizations from raw recordings of marmoset monkeys, resulting in a large vocalization dataset. This dataset will be the first milestone in future studies to train efficient computational models, such as DNNs, to learn high-level representations of monkey vocalizations. In humans, though, although we already have access to large datasets of voice samples (e.g., Common Voice dataset, Ardila et al., 2020), there are no existent neuroimaging datasets addressing the question of voice identity with sufficient data. To address this void, I conducted an extensive neuroimaging campaign to build a suitable dataset: numerous voice stimuli (10,000s) to leverage the link between the computational model and the brain activity in response to voice identity; stimuli balanced in several voice identity features: speaker's gender, age and identity; short stimuli duration (250 ms) to reduce the focus on speech content. In Chapter 3, I examine the correlation between brain activity evoked by voice identity and representations derived from deep learning.

2. Functional neuroimaging

Neuroimaging, or brain imaging, uses various techniques to visualize the central nervous system's structure, function, or pharmacology. Researchers use functional neuroimaging to explore how certain brain areas relate to specific mental functions. In these studies, participants carry out tasks while their brain activity is recorded using tools like electroencephalography (EEG), magnetoencephalography (MEG), or functional magnetic resonance imaging (fMRI). The collected data is then analyzed to identify patterns and correlations between specific brain activities and the tasks undertaken. fMRI technique is depicted in Figure 2.

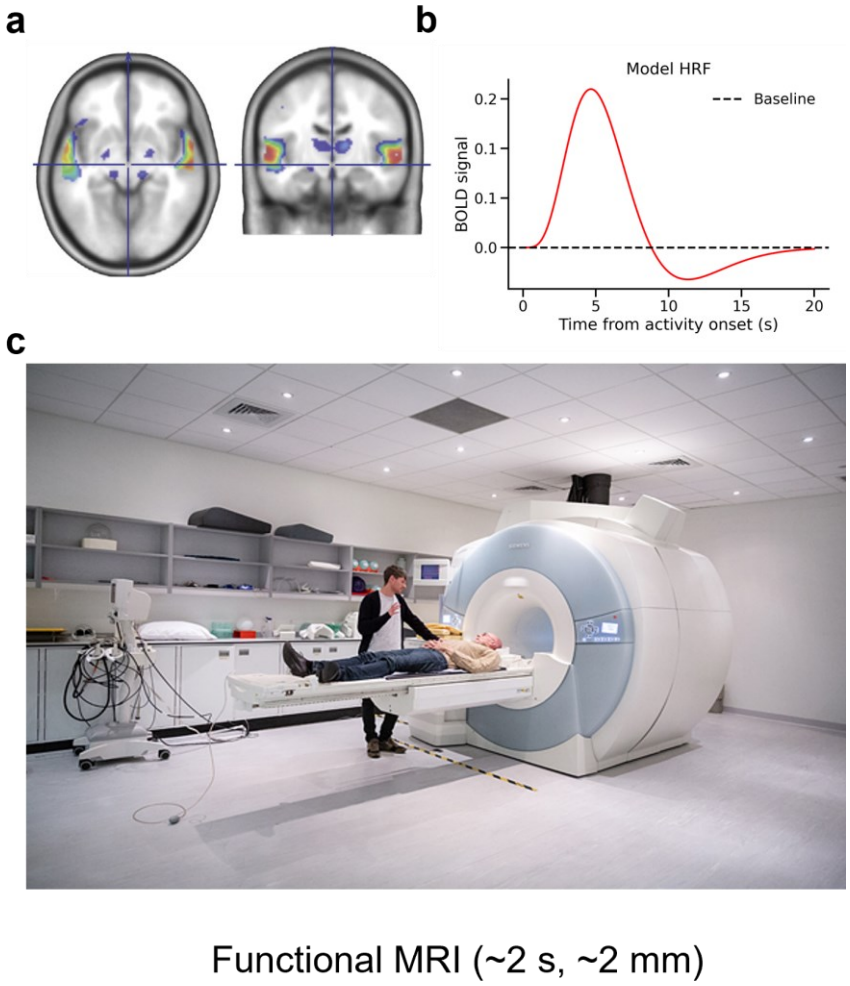


Figure 2: Neuroimaging technique used in this manuscript. a, fMRI maps of voice-selectivity highlighting the "temporal voice areas" (TVAs). Figure extracted from Pernet et al. (2015). **b,** Example of BOLD signal in response to one stimulus. **c,** A person about to undergo an MRI. Boxes around the ERF components of interest indicate the ±15 ms time window statistically tested (*P < 0.05; n. s., not significant). Figure extracted from Capilla et al. (2013).

Since its inception in the early 1990s, functional magnetic resonance imaging (fMRI) has emerged as a favored method for examining human brain function. This non-invasive technique does not necessitate injecting tracers or exposure to X-rays, making it suitable for a broad spectrum of participants, including children, who can undergo multiple scans if needed. While fMRI boasts a high spatial resolution (approximately 2 mm), its temporal resolution remains relatively low, around 2 seconds (Poldrack et al., 2011).

2.1. Blood flow and neuronal activity

When participants engage in particular tasks or are exposed to certain stimuli, the neurons responsible for processing these stimuli become active, necessitating an increased oxygen supply for neuronal activity. The fMRI captures this signal by monitoring changes in blood flow, termed the blood-oxygen-level-dependent (BOLD) signal. Though the neuronal activity initiated by a brief stimulus is fleeting (measured in milliseconds), the subsequent BOLD response, referred to as the hemodynamic response function (HRF), is more protracted. An illustration of HRF is presented in Figure 3. As depicted, the hemodynamic response takes roughly 5 seconds to reach its maximum. This peak is due to the surge of oxygen-rich blood, which heightens the local concentration of oxyhemoglobin. The magnetic attributes of oxyhemoglobin induce local field uniformities, culminating in an uptick in the T2*-weighted MRI signal. As blood circulation stabilizes, the hemodynamic response's pinnacle subsides, and a reduction in oxyhemoglobin concentration triggers a subsequent dip in the fMRI signal. Roughly 15-20 seconds post-stimulus, the hemodynamic response reverts to its initial state (Glover, 1999).

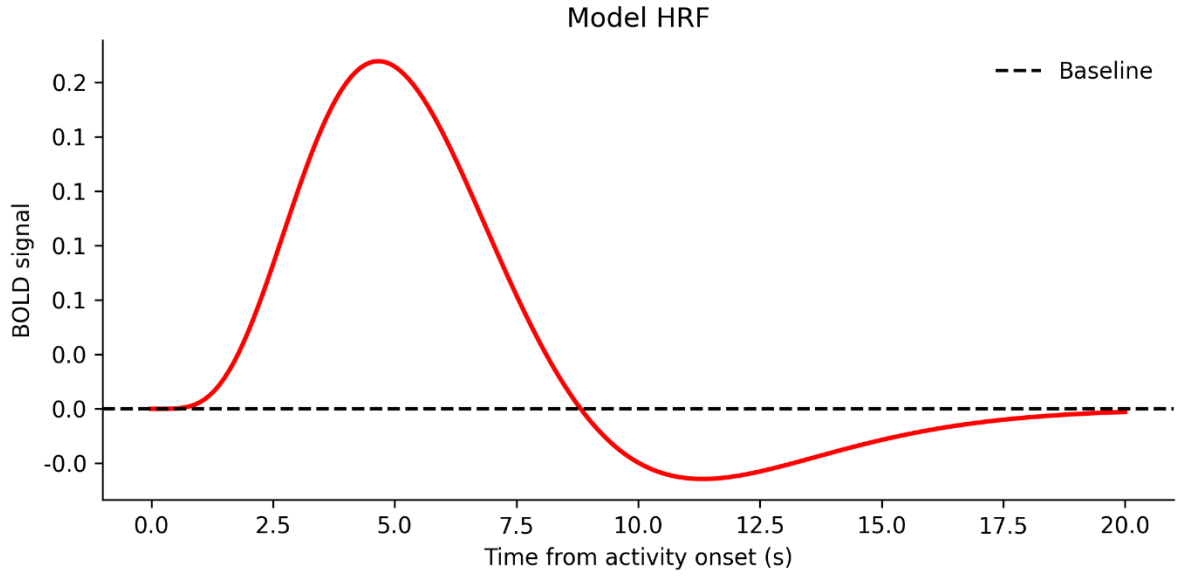


Figure 3: Hemodynamic response function.

2.2. Experiment designs for fMRI

During an fMRI experiment, participants are given a specific task, for instance, to localize the “temporal voice areas” (TVAs) within the human auditory cortex (Pernet et al., 2015). To achieve this, subjects passively listen to a sequence of both vocal and non-

vocal stimuli as their brain activity gets recorded. Various stimuli categories are typically presented to the participant at distinct times, with the fMRI data being captured concurrently. For example, if the repetition time (TR) is set at two seconds, an fMRI scan that reveals the current state of brain activity is procured every two seconds. To extract brain activity linked to a distinct stimulus, two primary experimental setups are examined in fMRI studies: block and event-related designs.

In the block design approach, stimuli are presented continuously for several seconds, which is then succeeded by a resting or baseline period of a similar duration. As the BOLD signal arising from a stimuli block is the aggregate of numerous individual responses, its amplitude is significantly larger than the signal produced by a singular, brief stimulus. An illustration of this block design can be seen in Figure 4. Such a design amplifies the nuanced differences between diverse experimental scenarios, making them more discernible. However, the BOLD signal derived from the block design represents a collective response of accumulated stimuli, offering limited insight into estimating the HRF associated with an individual stimulus.

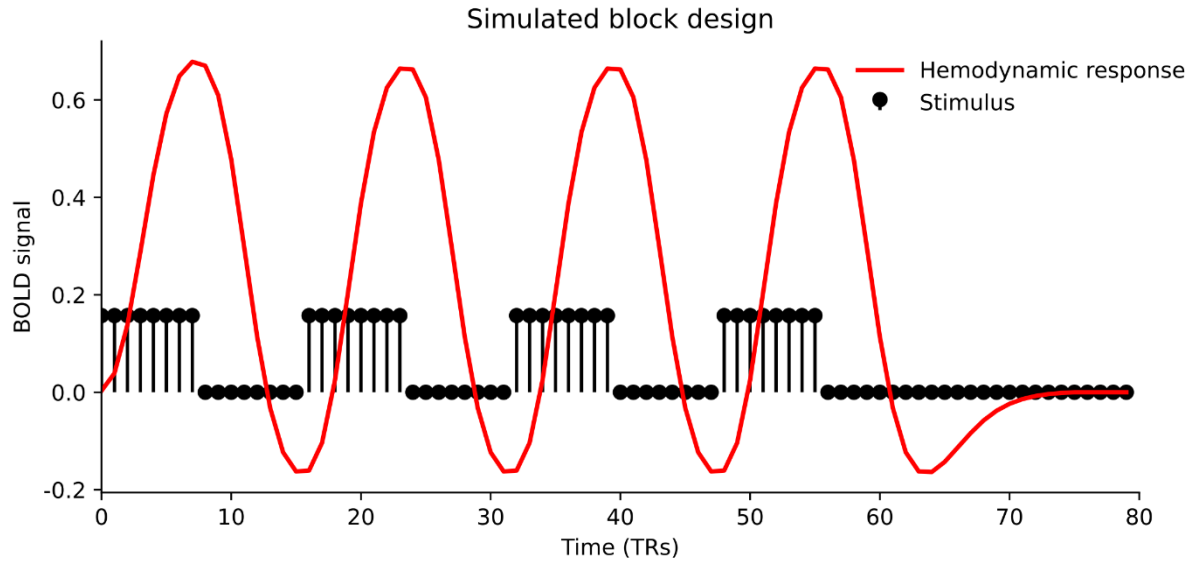


Figure 4: BOLD signal resulting from a block design. This simulation showcases a block design where a 20-second stimulus block is succeeded by a 20-second rest period. Notably, the amplitude of the BOLD signal in this design exceeds the amplitude of the HRF elicited by an individual stimulus.

An event-related design offers an alternative to the block design. In this setup, rather than showcasing several stimuli over an extended period, individual, short-duration

stimuli are presented with intervals in between, known as inter-stimulus intervals (ISI) (as depicted in Figure 5). The signals resulting from the event-related design typically have a smaller amplitude in comparison to the block design. Depending on the length of the ISI, event-related designs can be classified into either slow or rapid categories.

A slow event-related design utilizes an ISI that exceeds the HRF's duration, ensuring there is no overlap of individual hemodynamic responses. Considering the post-stimulus delay inherent to the BOLD signal, this design allows the hemodynamic response from a singular stimulus to peak and revert to the baseline. This design's advantage is that it facilitates estimating individual hemodynamic responses. However, its extended ISI can be seen as inefficient, leading to longer scanning times.

A quicker ISI is adopted to mitigate this inefficiency and fit more stimuli within a restricted timeframe, giving rise to the rapid event-related design. This framework sequences various stimuli in either a fixed or randomized order. The ISI is deliberately varied to yield a more consistent stimulus-response, meaning the time between stimuli is not constant. Due to the rapid event-related design's characteristic short ISI, the HRF duration sees overlapped individual responses within the BOLD signal.

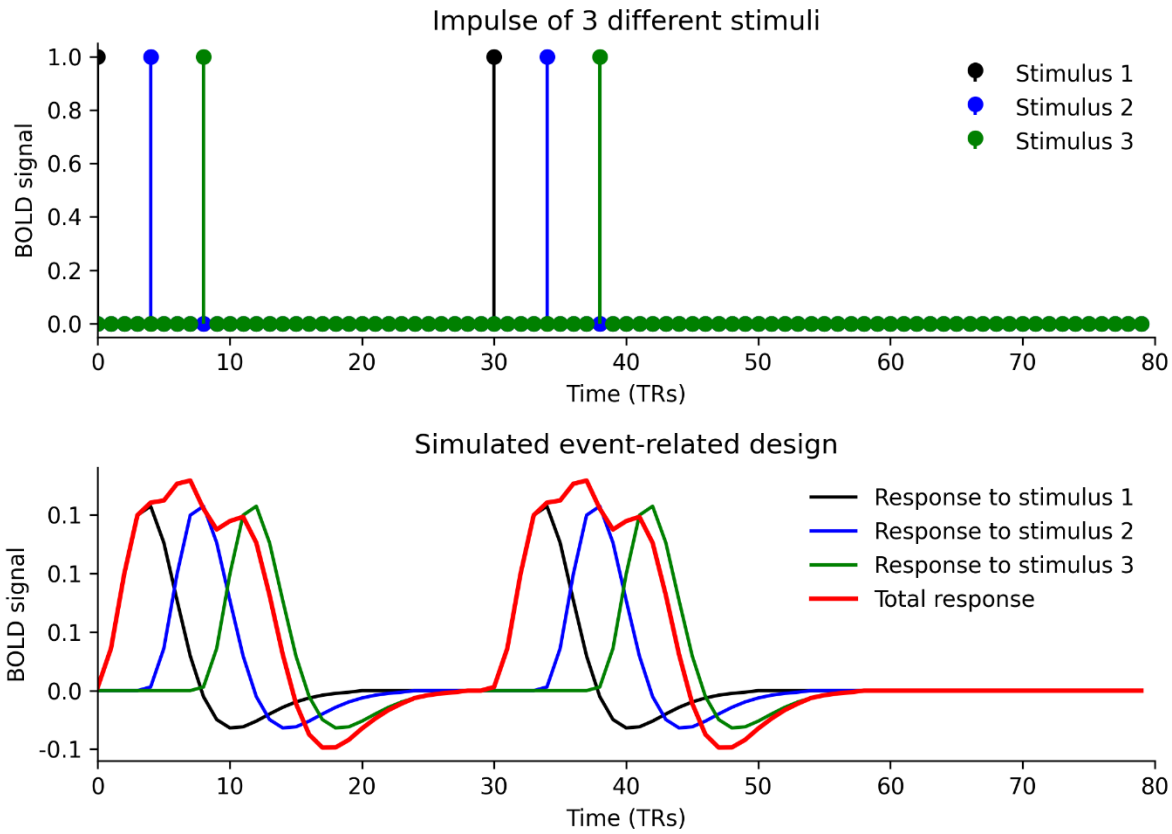


Figure 5: BOLD signal resulting from an event-related design. This simulated example presents three distinct stimuli, each separated by an inter-stimulus interval (ISI) of five seconds. Notably, the amplitude of the BOLD signal in this scenario is almost equivalent to the amplitude of the hemodynamic response function (HRF) elicited by a solitary stimulus.

2.3. fMRI data analysis

An fMRI dataset from an MRI scanner comprises a chronological series of three-dimensional volumes. Each volume comprises numerous small cubes, commonly known as voxels. Figure 6 illustrates fMRI volumes. To detect the specific brain functions linked to cognitive processes, e.g., pinpointing voxels associated with certain tasks, the responses from each experimental condition need to be estimated from the fMRI data.

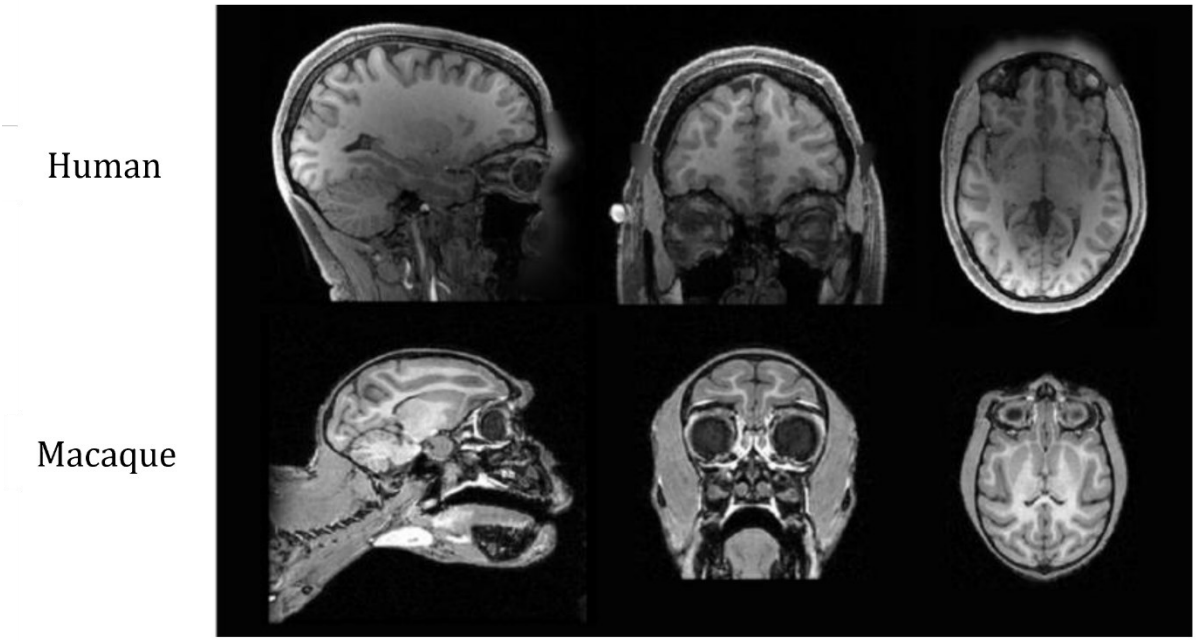


Figure 6: Example of 3D fMRI volumes. Schematic representation illustrating the differences in tissue structure between human and macaque brains as captured in MRI. Figure adapted from Wang et al. (2022).

Nonetheless, fMRI data can be contaminated with noise and may have various artifacts. A sequence of operations, commonly known as preprocessing, is applied to the fMRI data to address these issues. Typically, preprocessing of fMRI data encompasses the following steps (Poldrack et al., 2011):

- Quality control: MRI scanners can produce numerous artifacts. For example, electrical instabilities might result in spikes, while the heartbeat or respiratory patterns of the subject may lead to ghosting. Techniques like principal components analysis (PCA) and independent components analysis (ICA) are utilized to eliminate such artifacts and ensure data integrity (Bodin et al., 2021).
- Distortion correction: Echo-planar imaging (EPI) is frequently employed for fMRI data acquisition. However, magnetic field inconsistencies in EPI can introduce spatial distortions that may misalign subjects or displace activation sites. Methods, such as using magnetic field maps to determine the distortion extent, can help correct these distortions (Holland et al., 2010).
- Motion correction: Any head movements by participants during the scan can cause discrepancies in the position of the brain in consecutive images. This misalignment

can be mitigated by motion correction or realignment, where each fMRI image in the series is synchronized to a reference scan (Kim et al., 1999).

- Slice timing correction: Capturing an fMRI volume involves securing several two-dimensional slices and assembling them to create a three-dimensional structure. As these slices are captured sequentially, they have a time discrepancy. This discrepancy is managed by designating a slice as the reference and synchronizing the timings of the remaining slices to it (Sladky et al., 2011).
- Spatial normalization: Variances in individual brains pose challenges for population-wide brain function studies. To identify consistent patterns across participants, data from multiple subjects must be harmonized into a standard template, like the Montreal Neurological Institute (MNI) template (Cox & Hyde, 1997).
- Spatial smoothing: Enhancing the signal-to-noise ratio (SNR) is crucial, and to achieve this, high-frequency details are filtered out to diminish minuscule fluctuations in the image. Furthermore, spatial smoothing minimizes individual disparities (Mikl et al., 2008).

Once preprocessing is completed, the fMRI data exhibits reduced noise and an enhanced SNR. However, the intensity of the signal triggered by the task remains faint. For instance, task-activated voxels' percent signal change (PSC) typically ranges from 0.4% to 1% in block design. The PSC is even more subtle in event-related design, hovering around 0.1% (Mikl & Gajdos, 2014). Due to this, statistical models are employed to estimate the signal and evaluate differences across experimental conditions. Both univariate and multivariate approaches are utilized for data analysis from participants.

2.4. General linear model

The traditional statistical approach to analyzing fMRI data employs a univariate method. This technique operates independently on each voxel using the general linear model (GLM) (Friston et al., 1994). It is then applied iteratively across all brain regions to identify areas where the time-course is correlated with specific tasks.

Estimating the GLM parameters

The GLM is defined as follows:

$$Y = X\beta + \epsilon \quad (1)$$

where $Y = [y_1, \dots, y_N]^T$ denotes the BOLD signal time series for a specific voxel, representing the dependent variable with N observations at that particular location. The $N \times k$ design matrix, denoted as X, comprises k regressors; each serves as an explanatory variable; The vector $\beta = [\beta_1, \dots, \beta_k]^T$ is a column vector of k dimensions that needs to be estimated, corresponding to the i-th regressor of X. The error vector $\epsilon = [\epsilon_1, \dots, \epsilon_N]^T$ of size N, captures the discrepancy for each observation that not covered by the weighted sum of explanatory variables. Figure 7 provides a visual representation of the GLM.

To estimate the parameter β, the squared differences between Y and its estimate $\hat{Y}$ are minimized, $\hat{Y} = X\hat{\beta}$. $\hat{\beta}$ is obtained by

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2}$$

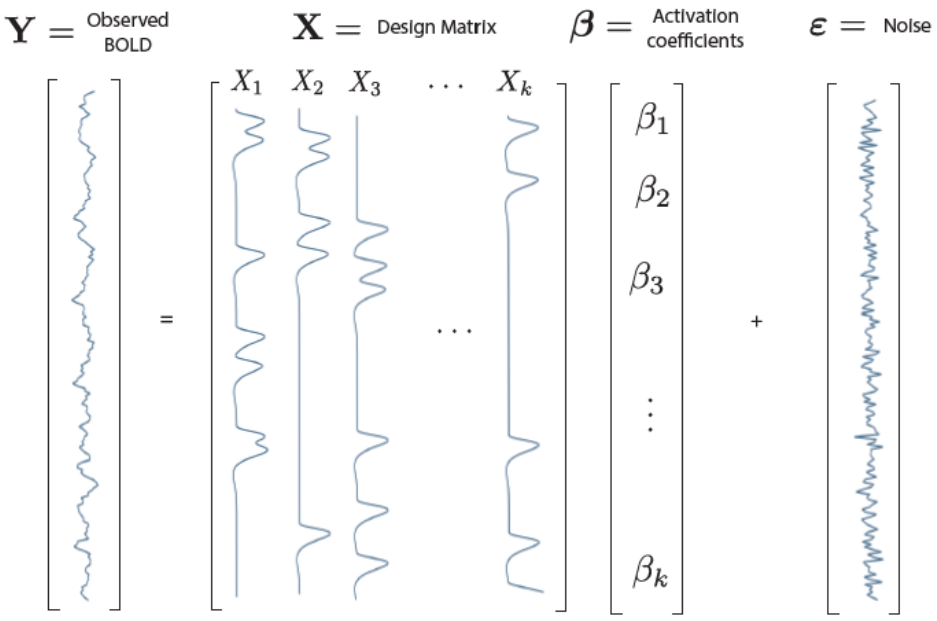


Figure 7: Illustration of the GLM. The GLM depicts the observed BOLD signal Y as a linear blend of regressors complemented by an error term (ϵ). Each regressor in the design matrix arises from convoluting a reference HRF with the stimulus function, set to 1 during stimulus presence and 0 otherwise (X). Every component of the undetermined activation coefficients signifies the relative magnitude of a specific condition (β). Figure reproduced from F. Pedregosa-Izquierd's thesis (2015).

Hypothesis testing

After estimating $\hat{\beta}$, hypothesis tests are carried out on contrasts (e.g. voice vs non-voice sounds for a voice localizer; Pernet et al., 2015). The null hypothesis is articulated as

H_0: $c\hat{\beta} = 0$, where c is either a vector or matrix of constants symbolizing one or multiple contrasts. As an example, if $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2]^T$ and the null hypothesis is expressed as H_0: $\hat{\beta}_1 = \hat{\beta}_2$, which can also be conveyed as H_0: $\hat{\beta}_1 - \hat{\beta}_2 = 0$, the contrast for testing if $\hat{\beta}_1$ differs from $\hat{\beta}_2$ will be $c = [1, -1]$. To verify the authenticity of the null hypothesis, a t-test is executed, yielding a t-value and its corresponding p-value.

Beyond just conducting a single t-test, it is possible to evaluate multiple contrasts using the F-test. For instance, when $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2]^T$, to concurrently test the null hypotheses H_0: $\hat{\beta}_1 = \hat{\beta}_2 = 0$, c would be represented as a matrix:

$$c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (3)$$

Following this, an F-test is conducted using c and the estimate $\hat{\beta}$ to deduce the statistic value and its p-value.

3. Evolution of primates

3.1. Primate phylogeny

3.1.1. Definition of primates and classification

Primates, derived from the Latin root "primas atis," are closely related to humans and boast a vast diversity, making them a prime focus for understanding the nuances of human evolution. Linné categorized them in 1758 as a part of the order of placental mammals. This group encompasses over 500 species found in various regions worldwide. Distinctive features of primates encompass an opposable thumb, flat nails, forward-facing eyes that grant stereoscopic vision, a limited number of teats, and a notable brain-to-body mass ratio.

Historically, primatology specialists employed the "Linnaean" method to divide the primate order into two sub-orders. This bifurcation labeled prosimians (comprising lemurs, lorises, and tarsiers) as the more rudimentary primates and the anthropoids (encompassing monkeys, great apes, and humans with larger brains) as the advanced group. However, this categorization has been critiqued for perpetuating a "species hierarchy" concept and lacking adaptability. Research indicates that tarsiiforms, a

subgroup within prosimians, share a closer kinship with the anthropoids, underscoring the call for a revamped classification system.

The adoption of the cladistic phylogenetic classification in the early 1970s brought about a shift towards the use of the phylogenetic tree for classification. Within this system, groups, or clades, are considered monophyletic: they encompass a theoretical ancestor and all of its subsequent descendants. Consequently, the primate order's taxonomy has been restructured to align, as closely as feasible, with these clades. This has led to a new division of the order into two suborders: the Strepsirrhinians (which were previously classified as prosimians, barring the tarsiers) and the Haplorrhinians (previously known as anthropoids).

Haplorrhinians, distinguished by the absence of a rhinarium (nose) and the presence of vibrissae (tactile whiskers), are further segmented into Tarsiiformes and Simiiformes. These further branch out into Platyrrhinians (or 'New World' monkeys, also called American monkeys) and Catarrhinians ('Old World' monkeys and hominoids, sometimes European and Asian monkeys). While Platyrrhines are recognized by their broad nostrils and elongated, prehensile tails, Catarrhines exhibit closely set nostrils that face downward and often do not have a tail. The graphical representation of this refined primate classification can be viewed in Figure 8.

3.1.2. Models in neuroscience

The predominant species utilized in neuroscience, collectively termed "Non-Human Primates," span both New World and Old World primate categories. Their varied morphological features, behavioral traits, and ecological niches enable researchers to select the most fitting model for their investigative pursuits. For instance, the common marmoset has been a staple in biomedical studies since the 1960s, while the macaque, due to its closer phylogenetic ties to humans, is frequently employed in foundational research. Neuroscientific exploration of great apes is infrequent, primarily because of logistical and ethical constraints; most research revolves around post-mortem anatomical analyses. Ethological investigations into these primates are indispensable, offering insights into human evolutionary trajectories. The macaque, with its thoroughly mapped anatomy and functions, stands out as the preferred model, particularly for neuroimaging techniques. Macaques belong to the old-world monkeys, sharing a common ancestor with

humans from about 28 million years ago (Steiper & Seiffert, 2012). However, the marmoset is seeing renewed attention, especially in auditory neuroscience, given its intricate vocal interactions (Miller et al., 2016; Okano et al., 2016). Marmosets belong to the platyrrhine lineage, sharing a common ancestor with humans from about 49 million years ago (Steiper & Seiffert, 2012), emphasizing their significance in studies (Figure 8).

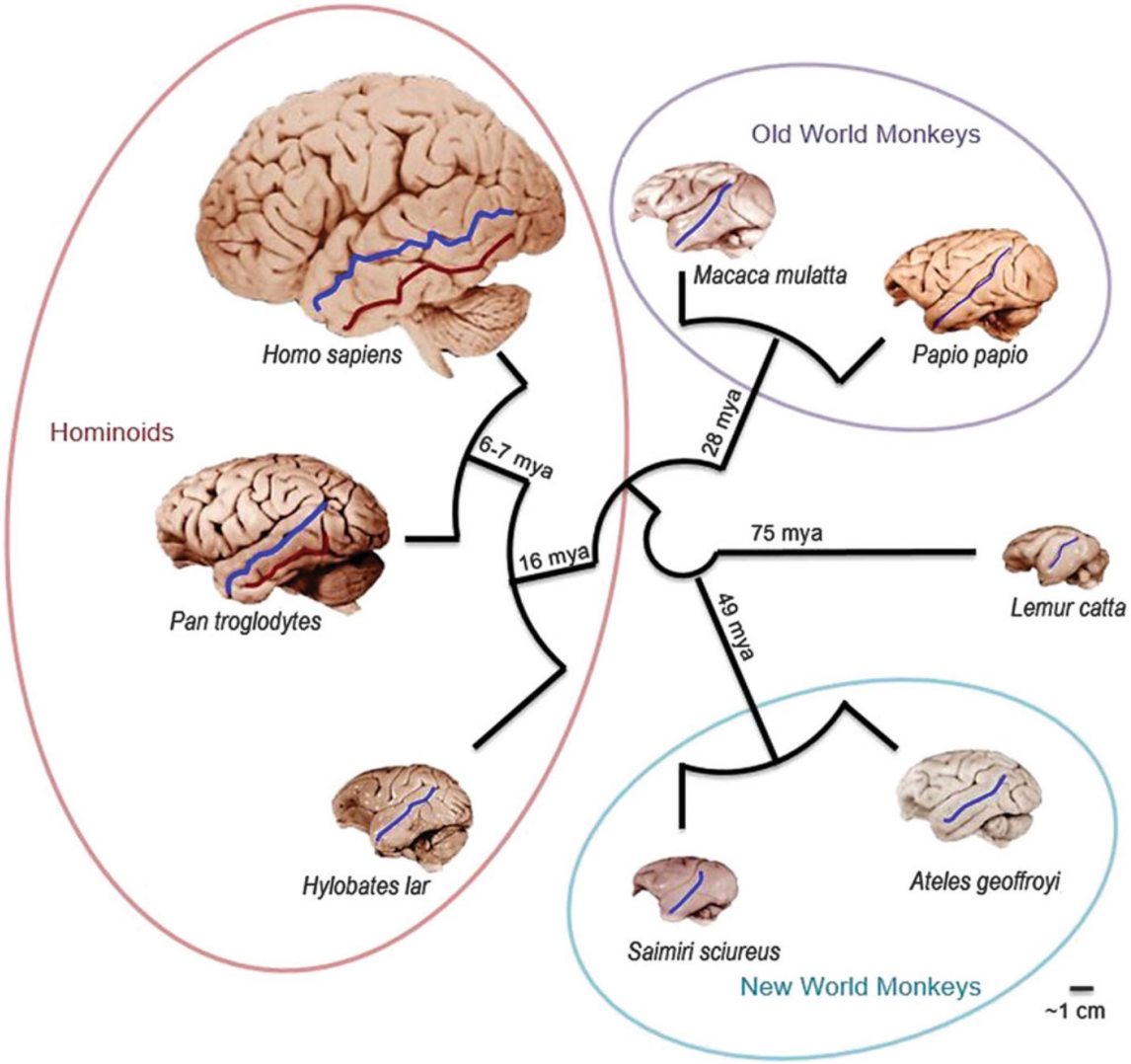


Figure 8: The evolutionary relationship of extant primates. Blue line highlights the superior temporal sulcus, present in the majority of primate species. Red line highlights the inferior temporal sulcus, present in hominids (chimpanzees and humans)**.** Divergence date estimates are from Steiper & Seiffert (2012). Figure reproduced from Bryant & Preuss (2018).

3.2. Auditory cortex

3.2.1. Evolution

Depicting primary cortical regions in mammals through a phylogenetic tree to underscore organizational homologies facilitates determining the hypothetical homology of their last common ancestor (Figure 9). This representation reveals that the broad spatial relationships among primary sensory areas remain consistent across mammalian species. In primates, these primary zones are particularly constrained in size, giving way to expansive higher-level associative regions (Krubitzer & Kahn, 2003; Buckner & Krienen, 2013). The primary auditory cortex in primates is situated within the temporal lobe, adjacent to the lateral sulcus. Of note, only this specific region has been identified across all the studied species, whereas secondary regions prove more elusive in their characterization and frequently bear varying nomenclature.

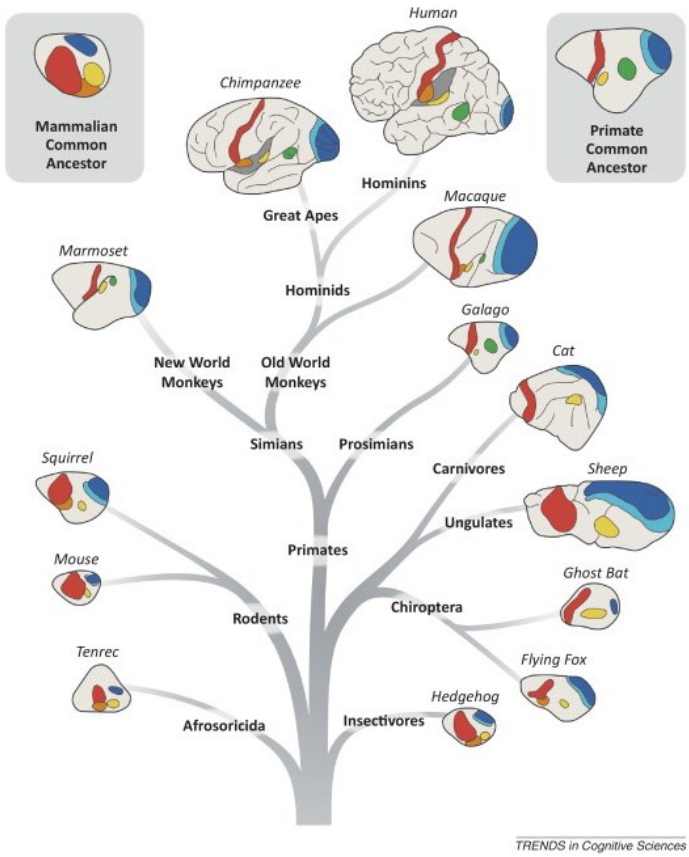


Figure 9: Phylogenetic representation of the cortex and primary sensorimotor areas. The primary auditory cortex is highlighted in yellow. Dark blue represents the primary visual area (V1); light blue indicates the secondary visual area (V2); green portrays the middle temporal area (MT); red signifies the primary somatosensory area (S1); and orange marks the secondary

somatosensory area (S2). The visual cortex is shaded in blue hues, while the sensorimotor cortex is depicted in red. The visual movement (VM) region is illustrated in green. Figure adapted from (Krubitzer & Kahn, 2003; Buckner & Krienen, 2013).

The extent of cortical myelination can serve as an indicator for pinpointing primary regions. Myelination is the process of forming a myelin sheath around select nerve fibers. This sheath, formed by the coiled wrapping of glial cells around the axon, facilitates faster nerve signal transmission. In Figure 10, areas with high myelination are marked in red for humans, chimpanzees, and macaques. The primary auditory cortex is visible along the lateral sulcus, though a segment is obscured behind the parietal operculum.

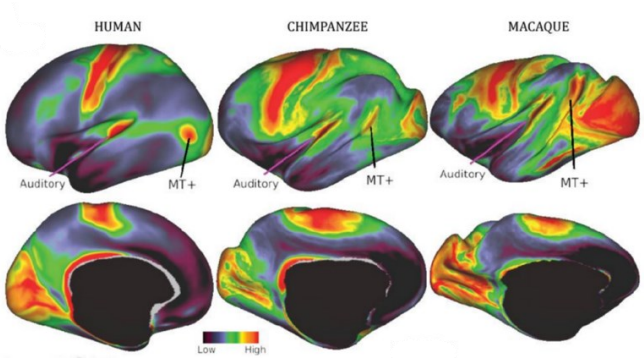


Figure 10: Localization of the primary auditory cortex in humans, chimpanzees, and macaques (maps depicted are not to scale). Myelination map illustrates the distinction between primary regions (with high myelination) and associative areas (with low myelination). Data provided by the Human Connectome project, WU-Minn Consortium.

3.2.2. Anatomy

Figure 11 presents a transverse section of the auditory cortex, offering a clearer view of its location and the medial-to-lateral gradient of primary and secondary subregions in three primate species: humans, chimpanzees, and macaques.

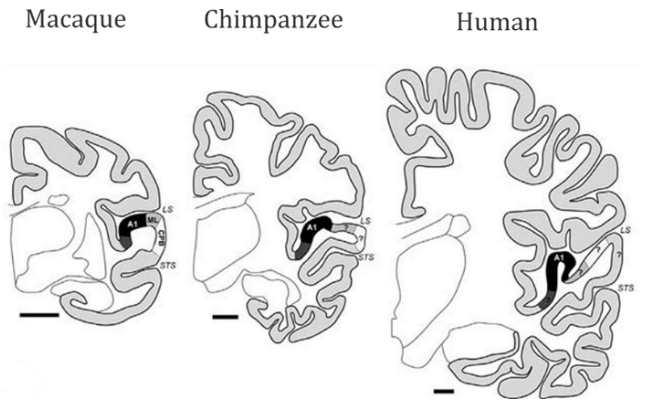


Figure 11: Auditory cortex across various primate species. A coronal illustrative view showcasing the location of the primary (A1 in black) and secondary (light grey and white) auditory regions. Scale bar 5 mm. Figure extracted from Hackett (2015).

In humans, the primary auditory cortex (A1) aligns with Brodmann's area 41, positioned between the Sylvian fissure and the superior temporal gyrus (STG). It is specifically believed to be situated along, and even matching the contour of, Heschl's gyrus (Da Costa et al., 2011). Multiple interpretations of the subdivisions of the human auditory cortex are available. Given the challenges of applying invasive tracing and electrophysiological methods in humans, this might account for the variances in such classifications. However, a general structural arrangement becomes evident, with the primary cortex (or core) encircled by the secondary auditory cortex—initially, the belt, followed by the parabelt stretching radially towards the extremities of the STG. It is essential to underscore that only the primary A1 region has been consistently identified across all primate lineages, with the subsequent regions necessitating further comparative research.

The hierarchical arrangement of the core, belt, and parabelt areas was first influenced by studies in non-human primates, especially the rhesus macaque, which serves as the primary model, but also research in the marmoset monkey (Eliades & Tsunada, 2019). Electrophysiological and fMRI examinations reveal that the medial and lateral belts of the marmoset auditory cortex house neurons responsive to vocalizations (Kajikawa et al., 2008; Rajan et al., 2013; Toarmino et al., 2017).

The auditory cortex (AC) in primates is structured in a hierarchical sequence of parallel fields. The primary core fields are encompassed by secondary belt and parabelt fields.

These secondary and tertiary regions process signals over extended durations, demonstrating increased sensitivity to specific intricate attributes and their combinations (Morel et al., 1993; Rauschecker, 1998; Hackett et al., 1998, 2007; Formisano et al., 2003; de la Mothe et al., 2006; Upadhyay et al., 2007; Bendor & Wang, 2008; Moerel et al., 2014; Cammoun et al., 2015; Schönwiesner et al., 2015; Tani et al., 2018; Besle et al., 2019). The core typically embodies two or three fields organized tonotopically, as depicted in Figure 12.

It remains ambiguous whether the A1 in humans aligns precisely with the A1 in other species (Ruthig & Schönwiesner, 2022). While cytoarchitectonic divisions have been identified in humans (Morosan et al., 2001), studies examining its structure have identified parallels with monkeys (Sweet et al., 2005; Fullerton & Pandya, 2007; Smiley et al., 2013). However, whether these regions share homology with non-human primate fields A1, R, and RT remains to be determined (Brewer & Barton, 2016; Besle et al., 2019). Active debates persist regarding the specific tonotopic map of the human auditory cortex and how it relates to the tonotopy in non-human primate auditory cortices (Baumann et al., 2013; Schönwiesner et al., 2015; Besle et al., 2019).

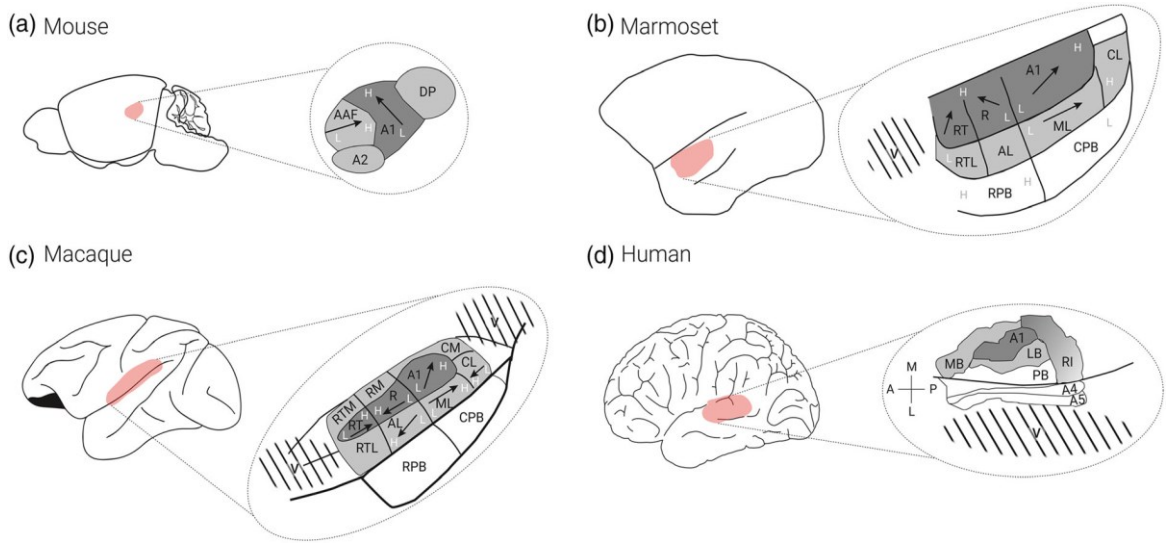


Figure 12: Schematic comparison of early auditory fields and adjacent voice-selective regions across various species. The auditory core, depicted in dark grey, and the surrounding belt fields, in light grey, exhibit a similar layout and tonotopic patterns in these mammals, though direct homologies remain unconfirmed. **(a)** Represents the mouse AC based on Stiebler et al. (1997). **(b)** Portrays the marmoset AC referencing Tani et al. (2018) and voice-selective zones from Sadagopan et al. (2015). **(c)** Illustrates the macaque AC per Hackett and colleagues (2001),

with voice-selective areas derived from Petkov et al. (2008, 2009). **(d)** Depicts human auditory fields as per Glasser et al. (2016) and voice-selective zones from Belin et al. (2000). Labels include A1, primary auditory cortex (termed auditory field 1 in humans); A2, secondary auditory cortex; AAF, anterior auditory field; AL, anterolateral belt; CL, caudolateral belt; CM, caudomedial belt; CPB, caudal parabelt; DP, dorsoposterior field; LB, lateral belt; MB, medial belt; ML, mediolateral belt; PB, parabelt; R, rostral field; RM, rostromedial belt; RPB, rostral parabelt; RT, rostrotemporal field; RTL, rostrotemporal lateral belt; RTM, rostrotemporomedial belt; V, voice selective areas. Figure reproduced from Ruthig & Schönwiesner (2022).

In summary, both in humans and monkeys, the auditory cortex structure reveals a functional hierarchy. Here, information primarily flows from the core region to more advanced regions (core > belt > parabelt > auditory related), moving from the lateral sulcus towards the ventral areas of the temporal lobe and then to associative regions beyond the temporal lobe. The discussed rostrocaudal connection gradient suggests a broader cortical division into dorsal and ventral streams for processing intricate sounds akin to what is observed in the visual system (Figure 13). The ventral stream, which links the rostral temporal lobe to the prefrontal cortex, likely plays a role in sound identification, whereas the dorsal stream seems to handle spatial localization and the sensorimotor representation of sounds (Kaas & Hackett, 2000; Rauschecker & Tian, 2000; Rauschecker & Scott, 2009; Rauschecker, 2012).

Balezeau et al. (2020) leverage a common MRI technique called diffusion-tensor imaging (DTI) to estimate the axonal (white matter) organization of the brains of macaques, chimpanzees, and humans. They echoed key findings from prior studies, in particular, a more dominant dorsal pathway – the arcuate fasciculus (AF) – in humans (Anwander et al., 2007; Rilling et al., 2012; Eichert et al., 2019), a more pronounced ventral pathway in chimpanzees (Rilling et al., 2012), a significant ventral pathway in macaques (Rilling et al., 2008), and a balanced ventral pathway (Figure 13, highlighted in green) observed across all three species (Rilling et al., 2008; Rilling et al., 2012). Insights into AF evolution drawn from functionally defined auditory regions show homologous ventral (depicted in dark green) and dorsal (in purple) pathways stemming from the AC in all three species. Notably, the AF segment appears left-lateralized in humans, a characteristic not as pronounced in nonhuman primates. These observations lend credence to the “primate auditory prototype hypothesis” proposed by the authors,

suggesting that the common ancestors of humans, apes, and monkeys may have had symmetrical dorsal pathways that linked auditory areas in the temporal lobes with the inferior frontal cortex (IFC).

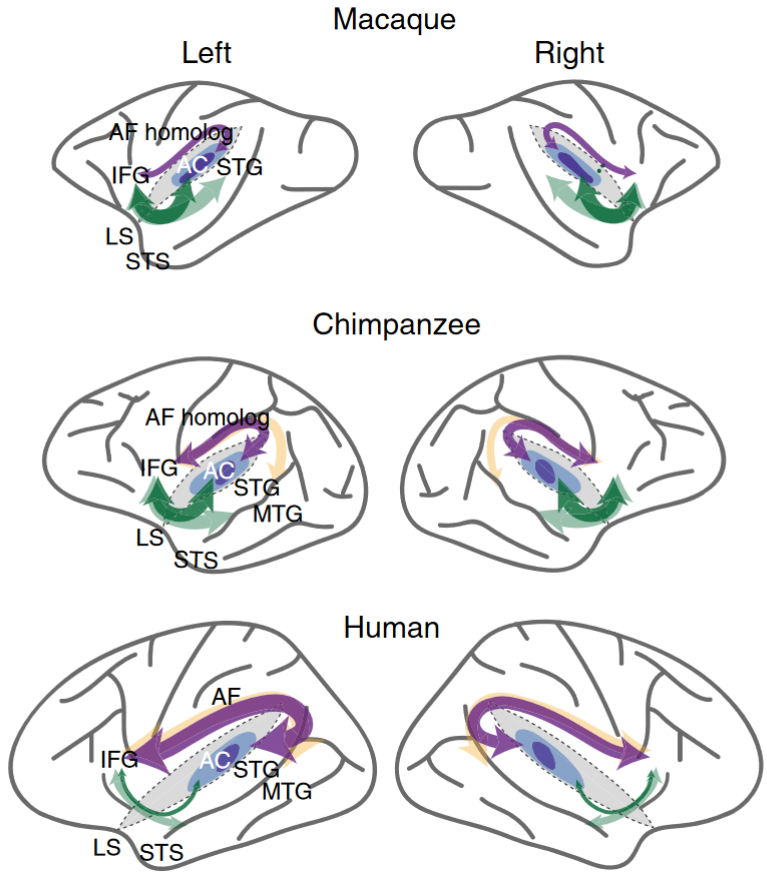


Figure 13: Comparative representation of auditory dorsal and ventral pathway strength and lateralization in macaques, chimpanzees, and humans. A visual summary of the dorsal (in purple) and ventral (in dark green) pathways for each species juxtaposed with earlier findings (represented in light yellow and light green). Figure extracted from Balezeau et al. (2020). Abbreviations: AF, arcuate fasciculus; IFG, inferior frontal gyrus; AC, auditory cortex; STG, superior temporal gyrus; LS, lateral sulcus; STS, superior temporal sulcus; MTG, middle temporal gyrus.

3.3. Vocal production

3.3.1. Source-filter theory

Vocal production encompasses actions executed by organs responsible for generating sound, including the lungs, larynx, nose, and mouth. The process initiates with generating a sound source, which subsequently experiences filtering by the organs specialized for this task. Initially recognized and detailed in humans, this mechanism has been termed

the "source-filter theory of voice production" (Fant, 1960; Fitch, 2000; Taylor & Reby, 2010).

The sound production mechanism is bifurcated into two primary phases (Figure 14). The initial phase is the "source", distinct for every individual, encompassing the larynx and all sub-laryngeal and laryngeal structures. The subsequent phase is the "filter", denoting the supra-laryngeal vocal tract that links the larynx to the mouth and nose openings, facilitating sound emission.

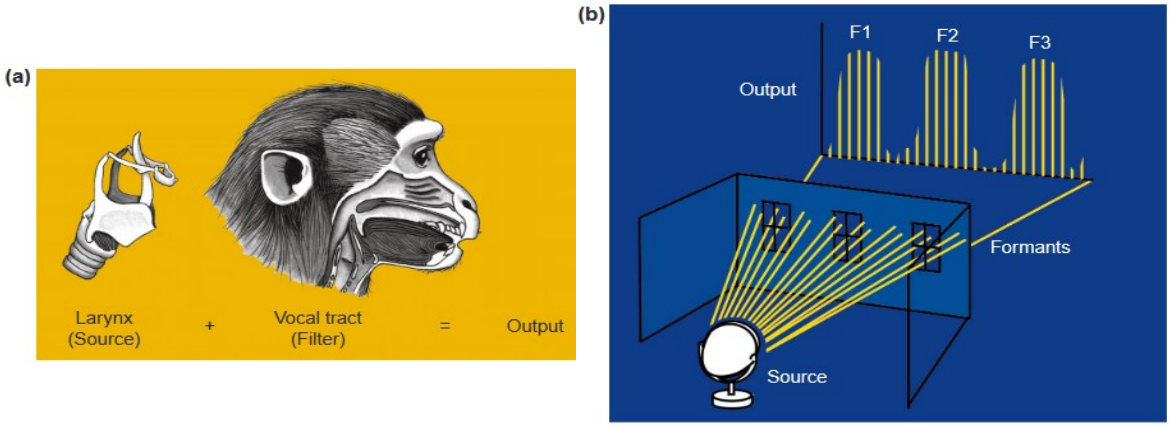


Figure 14: Illustration of the source-filter model of voice production. (a) Vocal sounds emerge from a sound source within the larynx, subsequently shaped by the vocal tract. **(b)** Formants arise from bandpass filters, operating as a frequency "window" to filter specific frequencies selectively. Figure reproduced from Fitch (2000).

During the initial source stage, the lungs provide the necessary airstream, and the larynx modulates this by governing the vocal fold movements. Anchored at the front by the thyroid cartilage and at the rear by the arytenoid cartilage, the vocal folds' actions are intricate (Fitch & Hauser, 1995). As the lungs expel air, biomechanical forces separate and converge the vocal folds. Such forces encompass the Bernoulli force, generated by the airstream moving between the folds, and the inherent elasticity of the folds themselves (van den Berg et al., 1957). The fluctuating air pressure in the larynx stems from the alternating opening and shutting of the folds. This movement rate sets the fundamental frequency, termed F0 (Fitch & Hauser, 1995; Taylor & Reby, 2010). The F0 is a pivotal acoustic metric influencing the pitch perception of a vocal sound. The source signal also yields other features, including rhythm, length, and volume.

The vocal tract, spanning from the larynx to the mouth and nasal cavity openings, is the site of the filtering process. Upon sound generation at the source, the vocal tract functions as a selective filter, either amplifying or attenuating specific frequencies from the source signal (Figure 14). These particular frequencies align with resonant frequencies, commonly termed formants. The final emitted signal is a composite of these formants (Fant, 1960).

3.3.2. Application of the source-filter theory to monkeys

While the "source-filter theory" was initially posited for human speech production, many studies have endorsed its relevance in explaining the vocal production of non-human primates (Rendall et al., 1998, 1999; Fitch, 2000). Research has shown that distinct acoustic structures of primate vocalizations are multifunctional. They not only distinguish between various call types but also relay information about the caller's identity, gender, and social affiliation (Rendall et al., 1998). For instance, one investigation highlighted the ability of macaque vocal tracts to uniquely alter the spectral structure of coo calls. Such modifications facilitate individual identification based on acoustic variations (Rendall et al., 1998).

Comparative anatomical examinations of larynx structures have revealed that both human infants and non-human primates possess a high-positioned larynx. This positioning suggests a shared inability to articulate until later in childhood when the larynx descends. However, it is widely believed that such descent does not occur in non-human primates (Negus, 1950). However, recent research has challenged the longstanding belief that larynx position directly correlates with vocal flexibility (Boë et al., 2017; Fitch et al., 2016). The "peripheral" hypothesis, predominantly propagated by Lieberman (Lieberman et al., 1972; Lieberman et al., 1969), is central to this discussion. This theory posits that the constrained nature of non-human primate vocalizations is a direct consequence of their vocal tract anatomy. Specifically, the pharynx cavity, which varies based on the larynx's vertical position, is thought to influence the diversity of achievable sounds significantly. Contrary to this, recent studies conducted on macaques (Fitch et al., 2016) and baboons (Boë et al., 2017) dispute the "peripheral" hypothesis, suggesting that, anatomically speaking, non-human primates have the potential to

produce a sound range comparable to humans. For example, marmosets possess a rich array of vocalizations, which will be explored further in the subsequent section.

3.3.3. Human vocal repertoire

Humans employ vocal communication as a primary mode of interaction, crucial for establishing and maintaining social relationships, exchanging information, and expressing emotions and intentions (Zhang et al., 2016). Their sophisticated communication system has evolved to cater to the multifaceted demands of social coordination, collective action, and cultural transmission (Smith et al., 2010).

Speech is the most recognized form of human vocalization, exhibiting a complex structure with varied rhythm, pitch, and timbre. This is organized into phonemes, morphemes, words, and sentences that convey different meanings in varied contexts. While speech can be broken down into numerous languages worldwide, its elemental features, like syllables and tonal variations, are universally present (Everett, 2005).

Besides speech, socially relevant information in voices, vocalizations, and voice perception is thus one of the significant sources of non-verbal and paraverbal auditory communication (Belin et al., 2004; Belin et al., 2011). For instance, laughter serves as a social bonding tool and a method to signify amusement or relief (Figure 15) (Scott, 2014). Crying, starting from infancy with basic hunger or discomfort signals, matures into a broader emotional expression spectrum in adulthood, from sorrow to joy (Figure 15) (Bell & Ainsworth, 1972). Furthermore, unlike uniquely human speech, these more basic channels of vocal communication are shared among many species (Nielsen & Rendall, 2018). This arguably makes the nonverbal channel of vocalization an even more powerful medium of social communication.

Singing is another distinct human vocalization, transcending mere speech by combining linguistic content with musical elements. It plays a pivotal role in cultural expression, religious rituals, and emotional self-expression (Welch et al., 1994).

Infants exhibit a unique set of vocalizations called babbling, which precedes speech. This babbling consists of repetitive, speech-like syllables and is a universal phenomenon, setting the foundation for later language development (Oller & Eilers, 1988).

Humans display an array of vocal modifications based on the situation and audience. For instance, in environments with increased background noise, humans instinctively raise their voices, known as the Lombard effect, to enhance speech intelligibility (Lane & Tranel, 1971). Humans also exhibit turn-taking in conversations (Sacks et al., 1978), with coordinated timing to avoid overlap and ensure a fluid exchange of ideas, echoing the antiphonal calls of marmosets (Miller et al., 2009).

The human vocal repertoire is diverse and adaptive, shaped by both evolutionary pressures and cultural nuances, enabling intricate communication within their complex social structures.

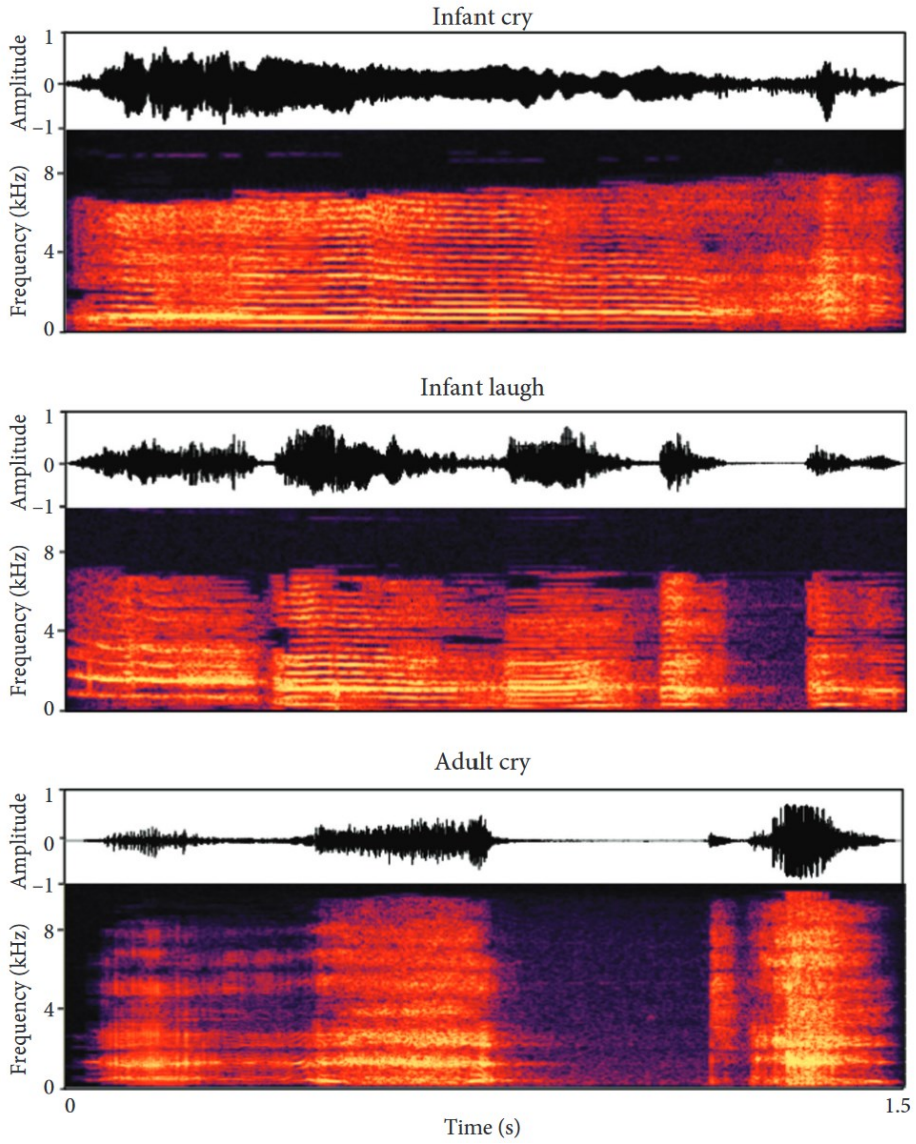


Figure 15: Examples of human voices. Waveforms and spectrograms demonstrating acoustic features of infant cry, infant laugh, and adult (female) cry sounds. While all types of vocalizations presented here have similar frequency ranges, the fundamental frequency (F0) of the infant cry (522 Hz) and infant laugh (562 Hz) is higher than that of the adult female cry (403 Hz). Patterns of burst duration also vary across sound types. Figure reproduced from Frühholz and Belin (2018).

3.3.4. Macaque vocal repertoire

In their natural environment, macaques utilize vocalizations to manage and harmonize group activities, employing various calls. Their vocal sounds can be grouped into 12 to 16 categories, varying according to the situation and emotional drive (see Figure 16 for examples) (Rowell et al., 1962; Hauser et al., 1991; Hauser et al., 1993). Field studies using the head-turning method to play back sounds have shown preferences in macaques for

conspecific (CV) and heterospecific vocalizations (HV) (Ghazanfar et al., 2001; Hauser et al., 1994), though these conclusions have faced some contention (Fitch et al., 2006; Teufel et al., 2010).

Early studies in Japanese macaques revealed their ability to discriminate various conspecific vocalizations (CV), especially from the "coo" class, more efficiently than other species (Beecher et al., 1979; Zoloth et al., 1979). However, later research suggested a nuanced transition between different CVs within the "coo" and "screams" classes (Rowell & Hinde, 1962; May, Moody, et Stebbins, 1988). The variability in each class may convey distinct information (Christison-Lagay et al., 2014). Macaques can discern identity from CV (Gouzoules et al., 1984; Hauser, 1991) and are sensitive to changes in formant frequency, which may relate to perceived body size (Fitch & Fritz, 2006; Fitch, 1997; Ghazanfar et al., 2007). While macaques exhibit a modest difference in body size between sexes, it is uncertain if they can discern gender based on conspecific vocalizations (CV).

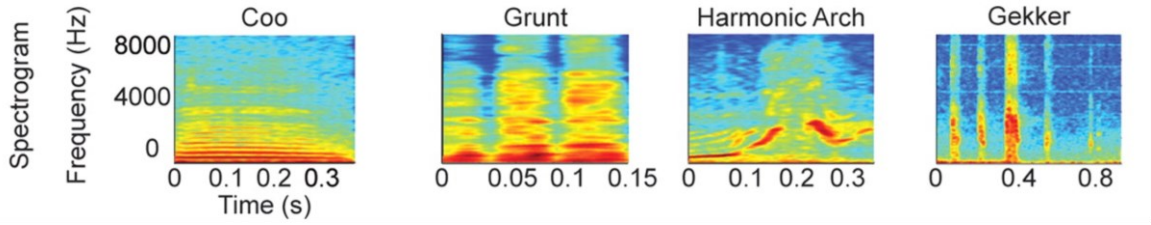


Figure 16: Examples of macaque vocalizations. Figure extracted from Averbeck and Romanski (2006).

3.3.5. Marmoset vocal repertoire

Marmosets, along with other primates, employ vocal communication in diverse situations, including predator evasion, self-defense, group travel, and food foraging (Tomasello & Zuberbühler, 2002). Additionally, they vocalize for territorial reasons and specific contexts tied to social interactions, play, and sustenance (Seyfarth & Cheney, 2003). Given the dense vegetation in which marmosets reside, their visual communication is restricted. They have honed a sophisticated vocal system to communicate effectively across distances and through visual barriers to overcome this.

Among the various calls, marmosets produce trills, and chirps are the most prevalent. Trills, characterized as extended whistled calls with sinusoidal frequency modulation,

function as intragroup contact calls (Bezerra & Souto, 2008). In contrast, chirps consist of short sequences of evenly spaced notes. Distinctively, the phee call is a prolonged intragroup contact call with a steady tone. It varies in form, such as the small phee, the long phee, and the loud shrill (Agamaite et al., 2015; Epple, 1968). The latter is emitted when marmosets are either separated from their group or marking territory (Bezerra & Souto, 2008; Miller & Wang, 2006). Another call, the twitter, is a succession of open-mouthed notes with ascending frequency, often used during encounters with other groups (Bezerra & Souto, 2008).

Additionally, marmosets have a range of alarm calls for atypical situations, encompassing sounds like tsiks, see or seep calls, screams, and chatters or cackles. They also generate non-melodic vocalizations, such as coughs, indicative of their anxious state.

Marmoset infants possess a unique vocalization known as the infant cry or nga, which evolves into mature sounds like phee call. Notably, marmosets can fuse vocalizations, creating combinations like cough-eks and trill-phees. Mirroring humans, they adjust their call's volume based on the perceived proximity of their audience (Choi et al., 2015) and tweak certain sonic properties in their reciprocal calls, termed antiphonal calls. These dialogic calls between members of the same species resemble human conversational patterns (Figure 17) (Miller et al., 2009).

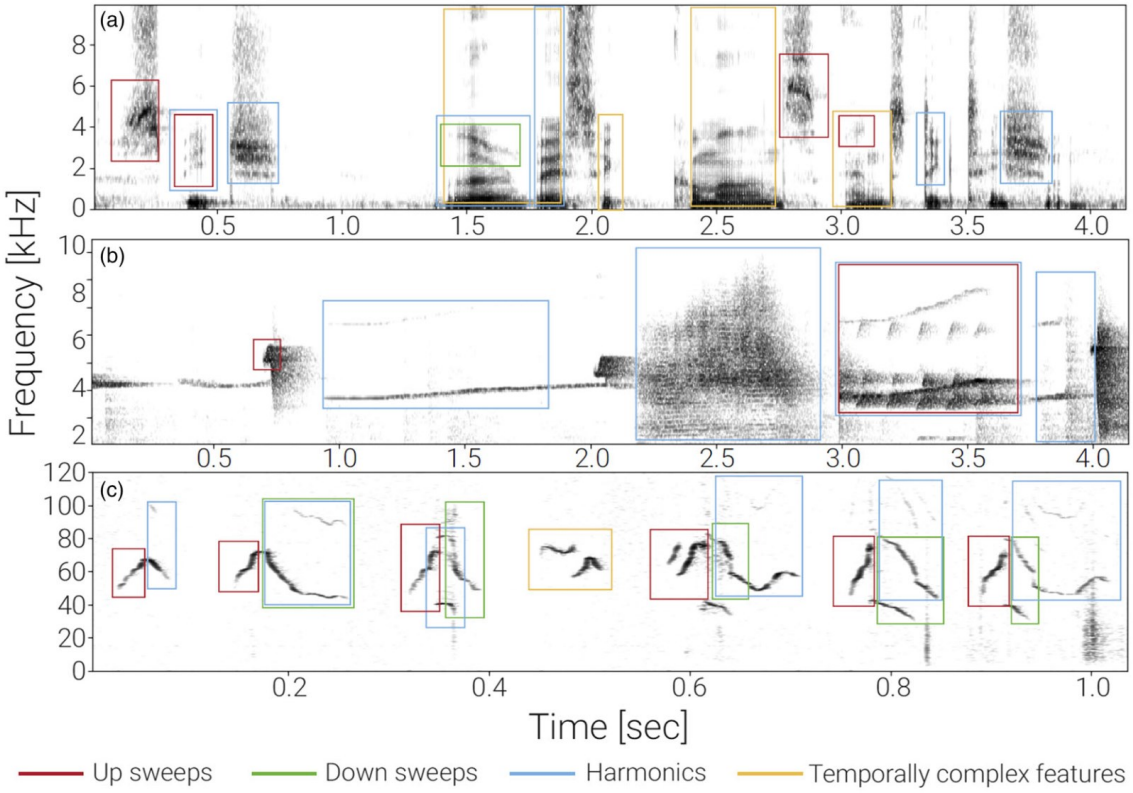


Figure 17: Example spectrograms of mammalian vocal communication. Spectrograms show acoustic features of (a) human speech, (b) a marmoset call, and (c) mouse ultrasonic vocalizations. All panels show longer vocalizations consisting of phonemes of varying acoustic complexity. Different acoustic components are highlighted (sweeps, harmonics, etc.). Figure reproduced from Ruthig & Schönwiesner (2022).

3.4. Vocal perception

Human and non-human primates utilize vocalizations to convey diverse information about external situations, such as threats, or to interact with others in various scenarios, including aggression, maternal actions, bonding exchanges, and beyond. Additionally, they heavily depend on these sounds to discern details about the vocalizer's identity, including aspects like gender, age, how well they are known, and other traits. Hence, precisely interpreting species-specific calls is vital for correctly understanding their social surroundings, even when visual indicators might be missing.

Perceiving vocal sounds has been a pivotal part of communication for numerous species long before the evolution of contemporary language. This positions it as a significant avenue to delve into the links between animal communications and the origins of human speech. Evidence suggests that humans and other primates share similarities in

perceiving voices, gleaning essential data from these sounds. Voice perception encompasses the ability to extract details from conspecific (from the same species) and heterospecific (from different species) vocalizations, including identifying the species, recognizing individual identities, and determining gender, along with interpreting the underlying emotions and intentions. While the ability to process spoken language is a human-specific trait, the perception of voice spans multiple species. Voice and spoken language are separate entities; the voice acts as the vessel transmitting speech details. Even without speech, the voice alone can relay much social and individual-specific information (Belin, 2018).

Humans have an exceptional ability to extract a broad range of information from vocal sounds (Belin et al., 2004; Belin et al., 2017). Through these auditory cues, one can identify speech, ascertain the identity, detect emotions, and even infer personality traits. As quoted, "We are all experts in voice" (Latinus et al., 2011). Research indicates that humans can effortlessly distinguish voices amid a sea of sounds. Even in brief sound intervals as short as 4 milliseconds, listeners have the aptitude to differentiate voices from other auditory stimuli, with their competence greatly exceeding mere guesswork (Suied et al., 2014). Intriguingly, this acumen in voice recognition does not transpose to other sound classifications within analogous durations; here, success rates merely hover around chance. When tasked to pinpoint specific auditory elements amidst distractors, individuals exhibit an amplified proficiency when the sought-after sounds are vocal (Isnard et al., 2016). This consistent pattern across diverse test conditions emphasizes voices' distinctive role in our auditory discernment.

This behavioral inclination towards the human voice finds its reflection in the neural pathways. The human auditory cortex, particularly the regions in the superior temporal gyrus (STG) and the superior temporal sulcus (STS), both anteriorly and posteriorly aligned with the primary auditory cortex, contains specialized zones termed "temporal voice areas" (TVA) (Belin et al., 2000; Belin et al., 2002; von Kriegstein et al., 2004; Pernet et al., 2015). These regions exhibit an augmented fMRI reaction to vocal stimuli, irrespective of their association with speech, in contrast to non-vocal auditory categories such as environmental sounds or heterospecific vocal utterances (HVs) (Belin et al., 2000; von Kriegstein et al., 2004; Fecteau et al., 2004; Agus et al., 2017).

The TVAs, the auditory equivalents to the visual cortex's "face areas" (Kanwisher et al., 1997; Haxby et al., 2000; Tsao et al., 2006; Freiwald et al., 2009; Hesse et Tsao, 2020), exhibit intricate structuring. Their precise anatomical positioning can fluctuate among individuals. However, certain research posits that they can be delineated by three "vocal patches" along the bilateral STG/STS: aTVA, mTVA, and pTVA (Pernet et al., 2015) (Figure 18a). Even though the overall activity in the TVAs is largely bilateral, some individuals manifest an asymmetry, with the right side of the temporal lobe displaying a heightened voice sensitivity in 33% of cases, compared to the left's 13% (Pernet et al., 2015). Additionally, the cerebral processing of voice extends beyond the auditory cortex, encompassing various prefrontal regions, notably the bilateral inferior frontal gyrus (Fecteau et al., 2005; Pernet et al., 2015; Aglieri et al., 2018) (Figure 18b).

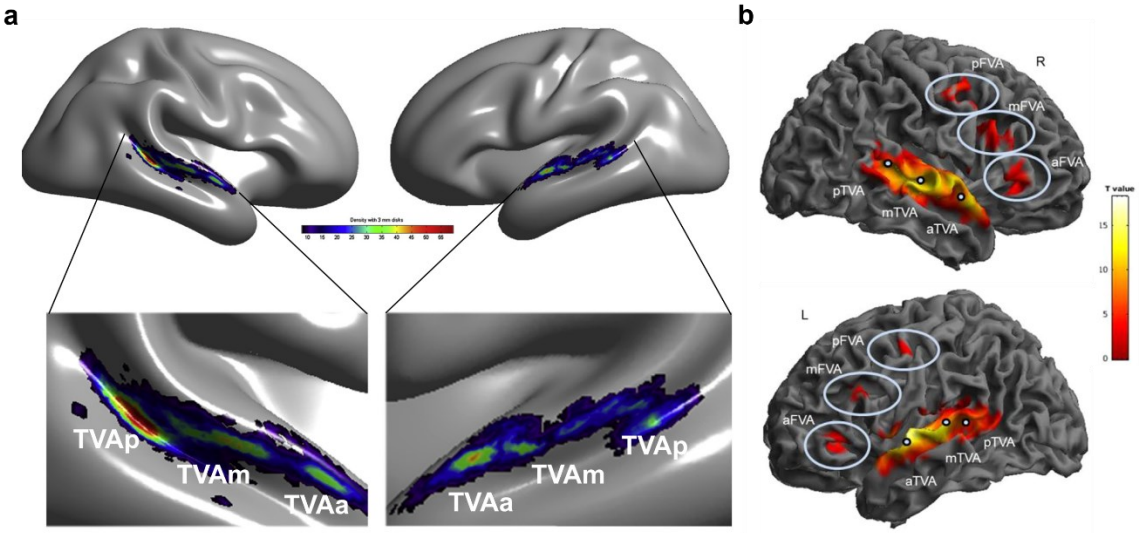


Figure 18: The human cerebral 'voice patches'. a, The TVAs in the human temporal lobe. Figure extracted from Pernet et al. (2015). **b,** The FVAs in the human frontal lobe. Figure extracted from Aglieri et al. (2018).

The auditory cortex of macaques has been extensively explored using a variety of techniques (Kaas et al., 1999; Rauschecker, 1998; Ghazanfar et al., 2004; Hackett, 2011; Romanski et al., 2009; Rauschecker et al., 2009; Ghazanfar et al., 2014). Electrophysiological recordings from awake animals indicate that neurons in the belt and parabelt areas of the secondary auditory cortex exhibit strong sensitivity to CVs (Tian et al., 2001), with increasing latencies and selectivity progressing in the caudo-rostral direction toward the temporal pole (Kikuchi et al., 2010; Fukushima et al., 2015). The

pronounced sensitivity of temporal lobe regions to CVs has been corroborated using whole-brain metabolic imaging techniques (Poremba et al., 2004; Gil-da-Costa et al., 2006;). With the advancement of fMRI in macaques, comprehensive cerebral estimates of CV sensitivity have been obtained using scanning protocols similar to those employed in humans. A macaque vocal area demonstrating responses analogous to human TVAs, i.e., favoring macaque CVs over other natural or control sound categories, has been identified (Petkov et al., 2008) (Figure 19). Employing fMRI-guided electrophysiology in the vocal area, it was shown that this region contains vocal cells, meaning individual neurons displaying vocal selectivity, akin to observations in facial patches of the visual cortex.

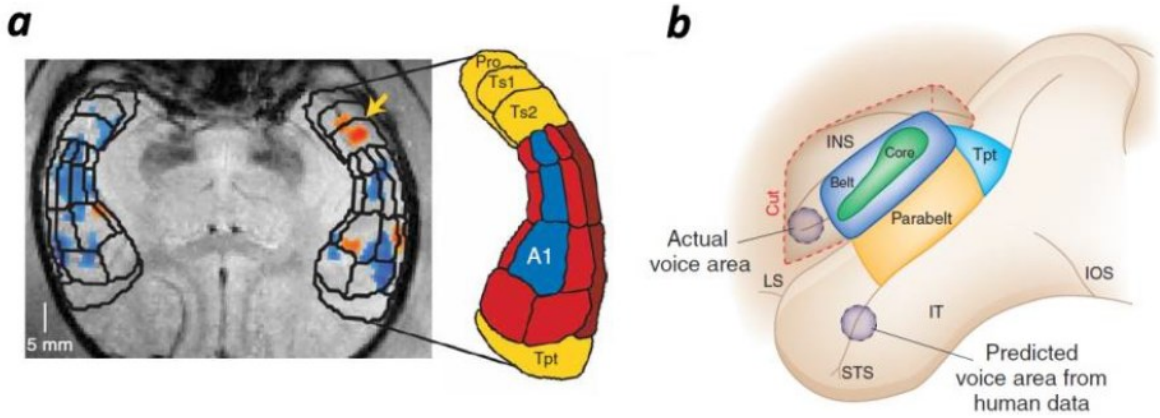


Figure 19: Vocal areas in the macaque brain. a, Functional MRI of the macaque reveals at least one vocal area (yellow arrow) with a preference for CVs in the anterior temporal lobe. **b,** The macaque's vocal area appears to be situated in a location not anticipated when drawing an analogy with human activation patterns.

A recent study by Bodin et al. (2021) unveiled a striking functional parallel between humans and macaques concerning the organization of the auditory cortex, especially within high-level areas dedicated to voice processing (Figure 20). Employing a uniform experimental procedure, humans and conscious macaques were examined through 3T fMRI scanning as they passively listened to an array of auditory stimuli, including human voices, macaque calls, marmoset calls, and other non-vocal sounds. The study found that both species possess voice-selective regions within the anterior temporal lobe that resonate specifically with vocalizations of their kind. Across 16 stimulus categories, A1 exhibited robust response patterns in both species. Notably, the correlation between hemispheres was particularly pronounced in humans, while it was barely significant in macaques.

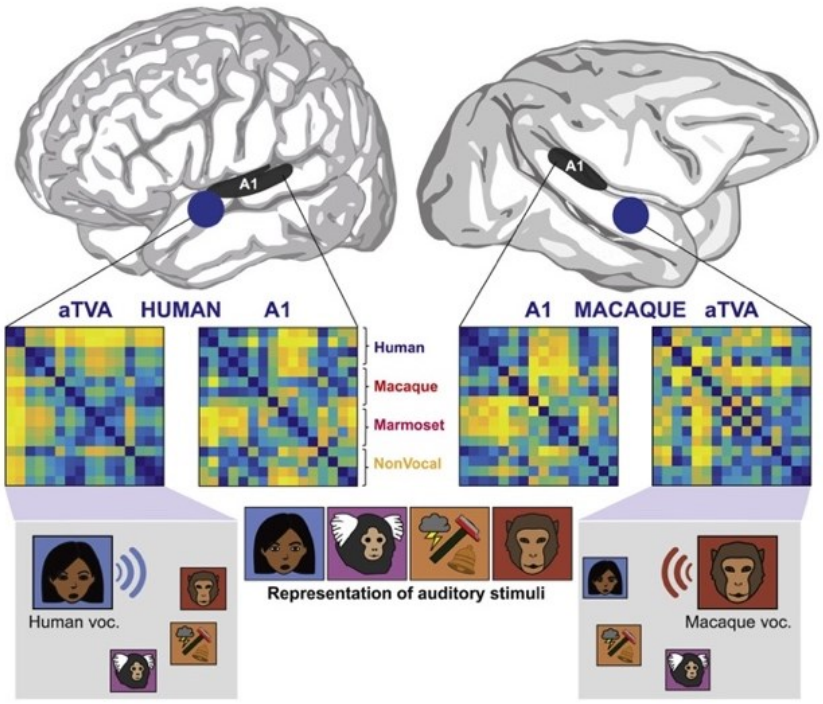


Figure 20: Functional homology in the cerebral processing of vocalizations by macaques and humans. fMRI data suggests that humans and macaques possess bilateral voice-specific regions in the anterior temporal lobe that favor conspecific vocalizations and process them similarly. Figure reproduced from Bodin et al. (2021).

Recently, marmosets have become increasingly popular subjects for neuroimaging studies. Their small stature allows for the use of high-field rodent MRI systems, providing enhanced signal and spatial precision. In a 2015 study involving six anesthetized marmosets, researchers explored the potential for voice-sensitive regions that prefer conspecific vocalizations. They utilized three types of stimuli: conspecific vocalizations (CVs), phase-scrambled CVs, and vocalizations from different animal species (Sadagopan et al., 2015).

To produce the phase-scrambled vocalizations, they derived the power spectrum from marmoset calls across six logarithmically spaced bands and randomized the phases of these bands. Subsequently, they merged them to create the final scrambled sounds. The data revealed that areas along the lateral sulcus, close to the temporal pole, displayed a particular affinity for CVs. Notably, the utmost rostro-lateral section demonstrated the strongest preference, as highlighted in Figure 21.

This pattern aligns with observations in macaques, hinting at a similar structure-function link in both species concerning the interpretation of conspecific vocalizations. This suggests that the cortical specialization for vocalization processing might have evolved roughly 40 million years ago in a shared ancestor.

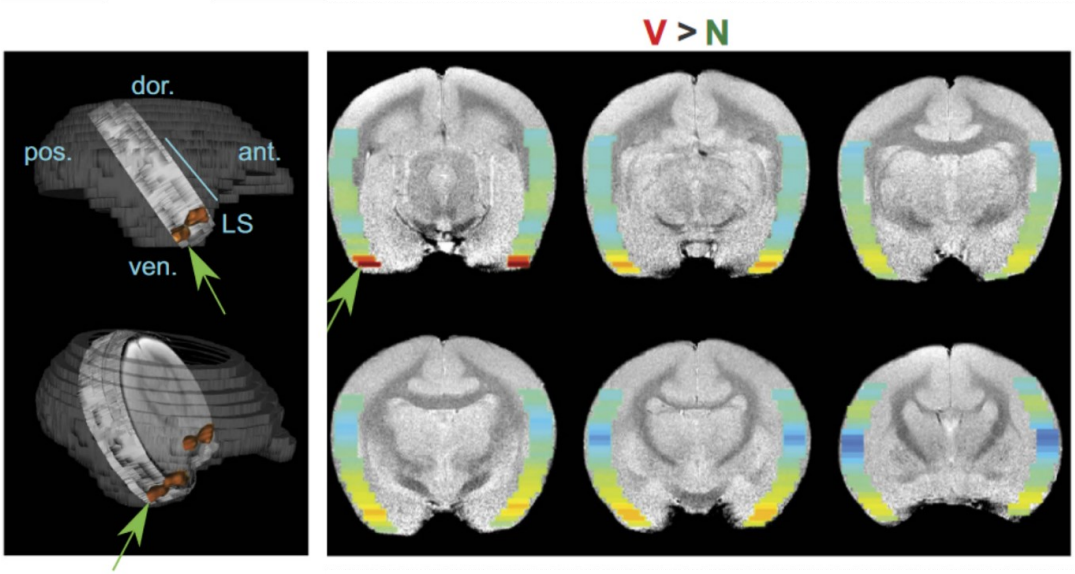


Figure 21: A caudal-rostral gradient for vocalization selectivity in the auditory cortex. The map displaying the differential response is projected back into anatomical coordinates, showing the gradient from caudal to rostral regions and indicating the location of the most selective area for vocalizations within the gradient. The green arrow and orange regions correspond to regions most selective for conspecific vocalizations. Figure reproduced from Sadagopan et al. (2015).

Another recent study by Stefan Everling, with ultrahigh field fMRI in awake marmosets, found a frontotemporal network, including subcortical regions, activated by conspecific vocalizations in marmosets (Jafari et al., 2023). They used three categories of auditory stimuli: CVs, time-scrambled CVs, and non-vocal sounds, including natural sounds, artificial sounds, and other animals. According to their findings in Figure 22, the activations did not show a caudal-rostral gradient (Figure 21) but rather at least 3 patches that may be homologs of the human voice patches.

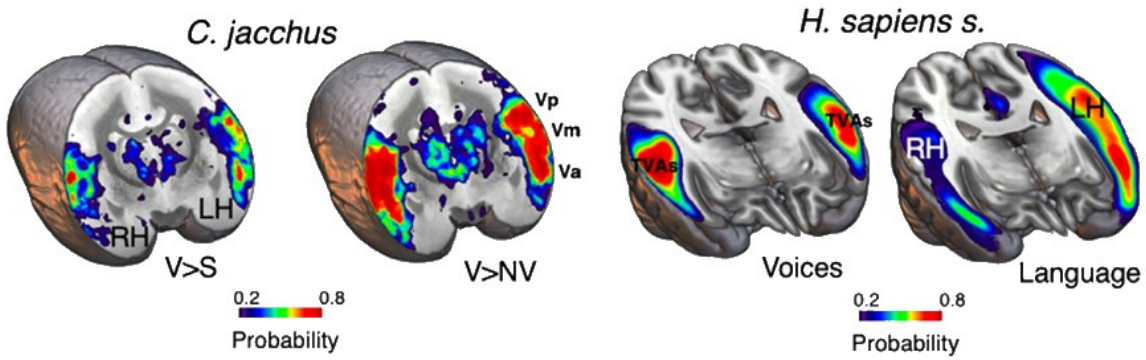


Figure 22: Comparison of marmoset and human networks. Left: Volumetric probabilistic functional map for vocal > scrambled vocalizations and vocal > nonvocal in marmosets overlaid on slices of anatomical MR images. Vp, vocal posterior, Vm, vocal medial, Va, vocal anterior. Right: Similar layout but for probabilistic functional human atlas for the voice and language localizer. TVAs, temporal voice areas. Figure extracted from Jafari et al. (2023).

These studies demonstrate a consistent functional organization of higher-level auditory cortex among various primate species. According to Belin and colleagues, this suggests the existence of a 'primate voice patch system' that specializes in processing conspecific vocalizations in primates (Belin et al., 2018; Bodin & Belin, 2020).

4. Representation learning with autoencoder-based models

In machine learning (ML), representation learning (RL) encompasses techniques that transform raw input signals into meaningful representations. When RL systems deploy multiple intermediate representations, it is termed deep learning (DL). RL is sometimes utilized purely for extracting features, and a subsequent machine learning system is employed for predictions, known as feature learning (Lee et al., 2009). In other scenarios, the RL system directly conducts inferences. Deep neural networks (DNNs) have recently surpassed other techniques in tasks like speech recognition (Hinton et al., 2012), visual object recognition (Krizhevsky et al., 2012), and natural language processing (Collobert & Weston, 2008), amplifying research and interest in this domain. Much of RL's foundation, particularly the principles derived from artificial neural networks (ANNs), drew inspiration from initial computational neuroscience models focused on neurons and their networks. Consequently, RL systems offer a more biologically grounded perspective than other machine learning systems (Bengio et al., 2014).

Autoencoder (AE) models in machine learning serve a dual purpose: firstly, they encode input data into a condensed latent vector, and subsequently, they decode it back to its original dimensionality. This process can be split into the encoder and the decoder, each offering many variations. In this section, I will delve into the diverse AE models we explored to extract a comprehensive yet compact representation from a dataset of vocal signals.

4.1. Principal component analysis

Principal component analysis (PCA) is a fundamental method for representation learning. Simply put, PCA determines the primary directions in which a dataset varies the most. It works by computing a linear transformation;

$$h = f(x) = W^T x + b \tag{4}$$

where x is the input data. Here, the columns of matrix W represent the main orthogonal directions of maximum variance in the dataset.

The new features or "principal components" are uncorrelated. This transformation allows the data to be represented in terms of these principal components. Less significant components—explaining the most minor variance—are typically discarded to reduce dimensionality. The resultant representation, often with fewer dimensions, can be more practical than the original data.

In Chapter 3 of our fMRI study, we established a baseline by investigating using PCA as a linear encoder to reduce the dimensionality of the input vector x. This approach was chosen because it has been demonstrated that a linear autoencoder with a d-dimensional hidden layer projects data in the same subspace as the one spanned by the d-first eigenvectors of a PCA (Plaut et al., 2018).

4.2. Autoencoder

Autoencoders (AEs), a specific category of Deep Neural Networks (DNNs), are formulated to learn a non-linear transformation that maps data from the signal domain into a reduced latent space during the encoding phase. Subsequently, they employ an inverse non-linear transformation through the decoding phase to reconstruct the latent

coefficients within their original domain (Vincent et al., 2010). This process is illustrated in Figure 23a.

Predominantly, they have been leveraged as an unsupervised method for reducing data dimensions. For instance, in the study by Hinton & Salakhutdinov (2006), AEs were tested on the grayscale images sourced from the Olivetti faces dataset. They juxtaposed the images reconstructed via AE with those achieved through the PCA, and they evaluated based on the same compression metric: the count of principal components in PCA versus the neuron count in AEs' bottleneck layer. Their findings highlighted that AEs substantially surpassed PCA for the datasets in question, generating images from latent descriptors that were profoundly analogous to the originals in quality and mean squared error (MSE) metrics (Figure 23b).

The encoder captures the input data x and maps it to a latent space representation z. The decoder then uses z to reconstruct the input, denoted as $\hat{x}$.

Given an input x, the encoder function, parameterized by weights ϕ, maps it to a latent space z:

$$z = f_{\phi}(x) \tag{5}$$

The decoder, parameterized by weights θ, then tries to generate $\hat{x}$ from z:

$$\hat{x} = g_{\theta}(z) \tag{6}$$

The training objective of an autoencoder is to adjust the parameters ϕ and θ to minimize the reconstruction error:

$$L(\phi, \theta, X) = \sum_{i=1}^{N} || \, x_i - \hat{x}_i \, ||^2 \tag{7}$$

where $x_i \in \{x_1, \ldots, x_N\}$ and N is the total number of training examples.

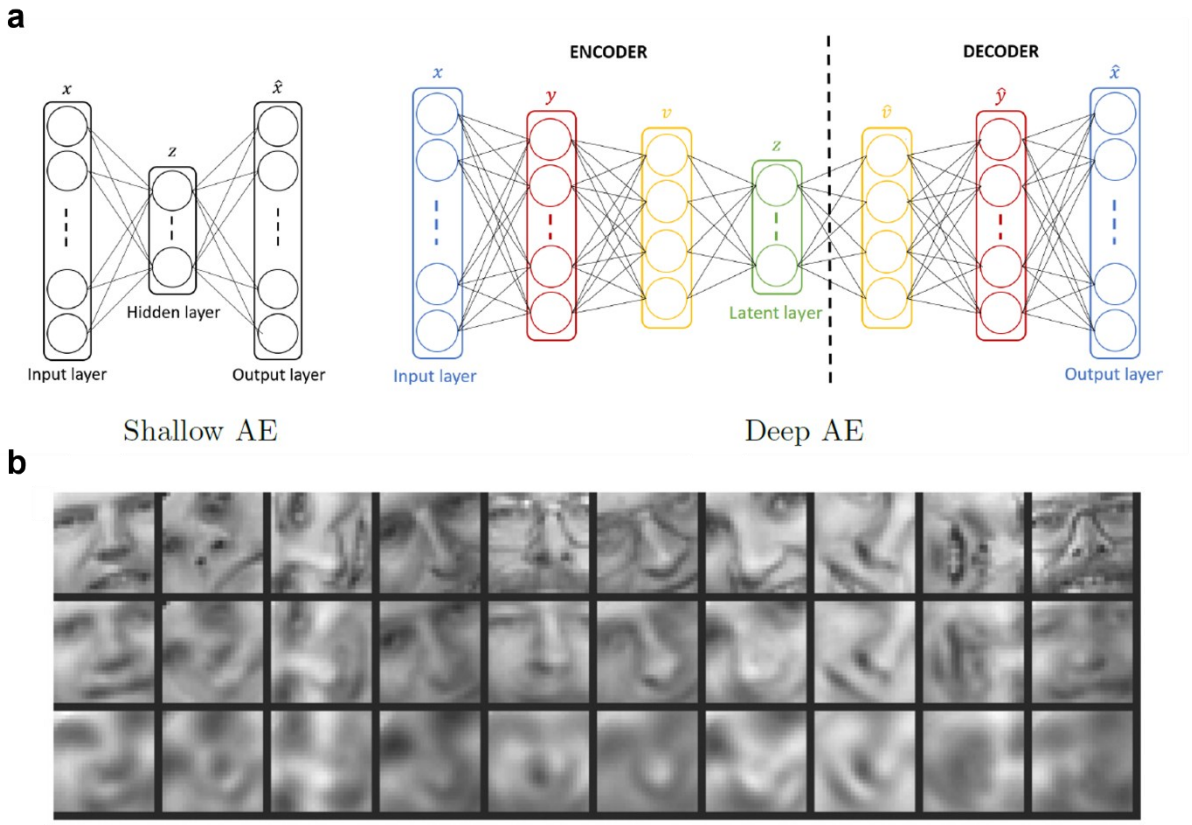


Figure 23: Shallow and deep AE. a, General architecture of shallow and deep autoencoders. The PCA can be seen as a shallow linear AE. Figure extracted from Fanny Roche's thesis (2023). **b,** Top to bottom: Random samples from the test data set; reconstructions by the autoencoder; reconstructions by the PCA. PCA gave much worse reconstructions. Figure extracted from Hinton & Salakhutdinov (2006).

4.3. Variational autoencoder

Variational autoencoder (VAE), introduced by Kingma and Welling (2014), represents a powerful approach in deep learning that uses neural networks for unsupervised representation learning from intricate data. Models of this nature have been widely employed for the synthesis of a diverse range of images: digits from the previously mentioned MNIST dataset (Kingma & Welling, 2014; Salimans et al., 2014), facial representations (Kingma & Welling, 2014; Rezende et al., 2014; Kulkarni et al., 2015; Higgins et al., 2017), compact images of tangible objects from the CIFAR dataset (Krizhevsky, 2009; Gregor et al., 2015), and even 3D renditions of chairs (Kulkarni et al., 2015; Higgins et al., 2017). They have also been used in forecasting subsequent sequences in static images (Walker et al., 2016). While VAEs are good at producing high-resolution images, they occasionally exhibit a mild blur. Their capacity to define a representation

space with notable characteristics, mainly due to the constraints imposed on the latent dimensions leading to a degree of decorrelation, positions them as promising tools for discerning valuable control parameters for synthesis.

Similar to their use in image synthesis, VAEs have recently gained traction in the audio domain. Initially, they were employed for the modeling, transformation, and synthesis of speech signals (Blaauw & Bonada, 2016; Hsu et al., 2017; Akuzawa et al., 2018). In a somewhat related context, VAEs have been utilized to model clean speech signals to enhance speech in noisy environments (Bando et al., 2018; Leglaive et al., 2018). Moreover, these models have applications in synthesizing musical sounds (Esling et al., 2018; Roche et al., 2021).

VAEs can be seen as a probabilistic/generative extension of standard AEs as, instead of deterministically mapping the input vector x to a unique latent vector z as done in AEs, the VAE encoder network maps x into the parameters of a conditional distribution $q_\phi(z|x)$ of z. Similarly, the decoder network maps a vector of latent coefficients z into the parameters of a conditional distribution $p_\theta(x|z)$ of x (Figure 24a). VAEs are thus considered as generative models as they try to capture the probability distribution of the data. Importantly, in a VAE, a prior can be placed on the distribution of the latent variables z so that they are well-suited for the control of the generation of new data, as exemplified by the speech interpolation in Figure 24b.

Given an input x, the encoder function parameterized by ϕ, ascertains the parameters (mean μ and variance σ^2) of the distribution over the latent variables z:

$$z \sim f_\phi(x) \tag{8}$$

Here, $f_\phi(x)$ is often modeled as a multivariate normal distribution using μ and σ^2.
The decoder, characterized by θ, reconstructs $\hat{x}$ from the sampled z:

$$\hat{x} = g_\theta(z) \tag{9}$$

The training objective for the VAE combines the reconstruction loss (difference between x and $\hat{x}$) and a regularization term:

$$L(\phi, \theta, X) = \sum_{i=1}^{N} (|| x_i - g_\theta\left(f_\phi(x_i)\right) ||^2 + D_{KL}(f_\phi(x_i) \,||\, p(z_i))) \tag{10}$$

Here, $x_i \in \{x_1, \ldots, x_N\}$; N is the total number of training examples; $D_{KL}(f_\phi(x) \,||\, p(z))$ represents the Kullback-Leibler divergence, which measures the difference between our latent space distribution and a standard normal distribution $p(z)$.

This loss function bears a clear interpretation. Since the KL divergence is always non-negative, $L(\phi, \theta, x)$ can be considered as a lower bound on the data likelihood (Doersch, 2021). This is often called the "Evidence Lower Bound" or ELBO.

For training VAEs, one can use gradient-based optimization techniques, especially when $f_\phi(x)$ represents a multivariate normal distribution with parameters μ and σ^2 (Kingma & Welling, 2014). The Adam optimizer, proposed by Kingma and Ba (2015), is a popular choice in this context.

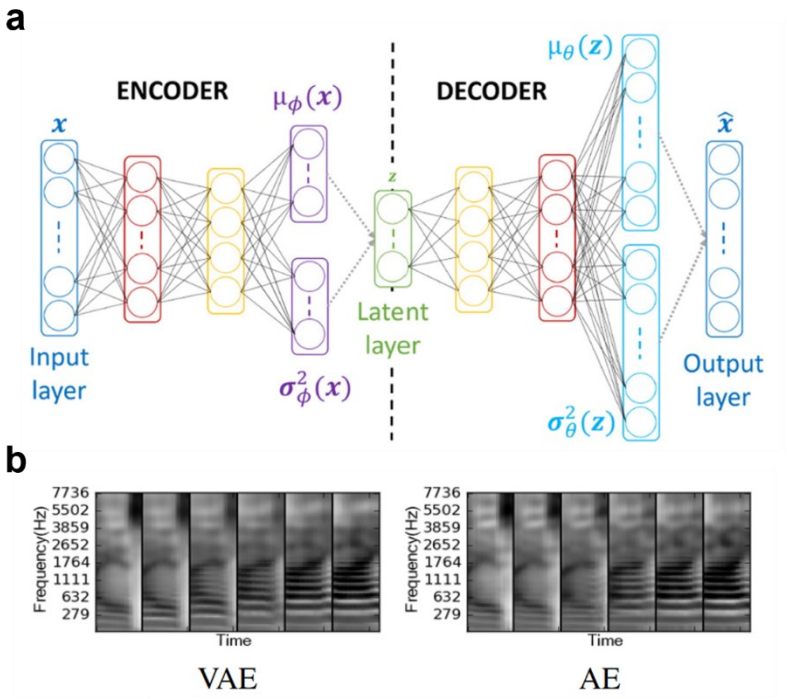


Figure 24: Variational autoencoder. **a,** VAE's general architecture with grey dotted arrows denoting the sampling process. Figure reproduced from Roche et al. (2021). **b,** 200 ms segment interpolation from a male /ey/ to a female /ay/ using both VAE and AE. The VAE transition highlights pitch and formant contour changes, whereas AE demonstrates a more direct feature space interpolation. Figure extracted from Hsu et al. (2017).

Chapter 1

Comparative study of the vocal cortex in primates

Voices play a pivotal role in communication among humans and other primates, encoding essential information such as gender, identity, and emotion. The ability to interpret these vocal cues is supported by a specialized brain system comprising interconnected cortical areas. These regions work together to form increasingly abstract representations of vocal sounds. Despite recognizing the significance of this system, our understanding of the specific contributions of each component still needs to be completed. This chapter aims to compile recent discoveries regarding the structure and function of these cortical areas in primate brains. By synthesizing these findings, we aim to construct a refined model of vocal processing, highlighting the distinct roles played by various brain regions in interpreting vocal signals.

1. Abstract

It has been suggested that the primate brain processes vocal information through a "voice patch system" (Belin et al., 2018) consisting of discrete, interconnected cortical areas supporting increasingly abstract vocal input representations. In a subsequent review, Bodin and Belin (2020) provided further evidence supporting the cerebral basis of conspecific voice (CV) in human and non-human primates. Their findings highlighted a conserved voice patch system in the temporal lobe, called the "temporal voice areas" (TVAs). Additional neuroimaging studies have identified extra-temporal regions with varying degrees of sensitivity to CV. Three bilateral regions in the human frontal cortex have been labeled as the "frontal voice areas" (FVAs) (Aglieri et al., 2018), showing greater sensitivity to vocal compared to non-CV stimuli. While it is clear that vocal information passes through these voice areas, the specific functions of each area are not yet fully understood. This chapter is dedicated to exploring the current understanding of this intricate system and identifying the gaps in our knowledge of it.

2. Understanding voice perception

Voices, crucial for communication in various species, especially primates (human and non-human primates), necessitate focused research on mammalian vocal patterns and the neural bases in voice-sensitive areas, known as voice areas (VA). These regions are defined by their higher responses to voices than to other auditory stimuli using functional brain-imaging studies. These studies indicate that distinct cortical brain regions show robust responses to voices after basic sensory processing in the primary auditory cortex (A1). While the processing of basic sound features in primates is well-established, there still needs to be more understanding regarding how the brain, especially within the temporal and frontal VA, converts intricate vocal signals into meaningful high-level representations. We present below a series of questions aimed at evaluating the state of knowledge within both recent and past literature to shed light on the limitations of the current literature in the field:

- ***What is the functional role of each unit within the "voice patch" system in the human brain when processing vocal information?***
- ***How do voice patches connect within the brain, what are their processing stages, and are there distinct temporal dynamics in voice processing?***

- ***How is voice identity encoded in the brain?***
- ***Is there a shared voice coding principle across primate species?***
- ***What are the voice recognition mechanisms?***
- ***Which computational models align with the VA representations?***

This review chapter first overviews current evidence, focusing on human anatomical, functional, and neural aspects. Next, we examine voice processing in other primates to find common patterns or homologies. Finally, we investigate the principles behind voice recognition, explore how computational modeling can reveal these principles, and discuss why combining these insights is essential for a unified understanding of voice perception and processing.

3. Anatomical organization of the voice processing system

The anatomical organization of the auditory cortex is thought to reflect a functional hierarchy where information mainly flows from primary regions to more secondary regions, along the superior temporal gyrus (STG) and sulcus (STS), and all the way to extra-temporal associative areas. Figure 1.1a illustrates the brain regions in both the left and right hemispheres involved in decoding semantic, identity, and emotion-related information from vocal cues. Regions are color-coded, with lighter shades indicating areas likely involved in post-perceptual processing —integrating initial sensory input with memories or broader knowledge.

The temporal voice areas (TVAs) are arranged bilaterally along the Superior Temporal Sulcus and Gyrus (STS/STG). The frontal voice areas (FVAs) are primarily located within the inferior frontal gyrus, spanning regions from the pars orbitalis to the junction of the precentral and middle frontal gyrus. The limbic system is an aggregation of brain structures generally located lateral to the thalamus, underneath the cerebral cortex, and above the brainstem.

A portion of the inter-individual variability in the occurrence and position of the TVAs/FVAs could be related to the high inter-individual variability in the anatomy of sulci patterns. Indeed, correspondence between the location of functional activations and

anatomy has been observed in multiple cases, such as for the location of the TVAs relative to the superior temporal asymmetrical pit of the superior temporal sulcus (Bodin et al., 2018) and for the location of the FVAs relative to the individual sulcal anatomy of the prefrontal cortex (Cordeau et al., 2023).

Temporal Lole The anterior temporal pole (aTP) is situated in the most anterior part of the temporal lobe (BA 38). The anterior temporal voice area (aTVA) is in the anterior superior temporal sulcus. On the left hemisphere, LaTVA is found in the superior temporal gyrus (STG) according to the Anat toolbox and as the STG anterior division in the Oxford atlas. On the right hemisphere, RaTVA is identified in the temporal pole using the Anat toolbox and is also labeled as the STG anterior division in the Oxford atlas. The left (LmTVA) and right (RmTVA) mid-temporal voice areas, as defined by Pernet et al. (2015), are located in the middle superior temporal gyrus and sulcus/gyrus, respectively. Both are identified in the superior temporal gyrus (STG) by the Anat toolbox and labeled as the STG posterior division in the Oxford atlas. The posterior temporal voice area (pTVA) plays a crucial role in auditory-motor integration. Anatomically, both the left (LpTVA) and right (RpTVA) sections are identified in the middle/posterior superior temporal gyrus by Pernet et al. (2015). The Anat toolbox places them in the superior temporal gyrus (STG), with the Oxford atlas categorizing both as the STG posterior division.

Frontal Lobe The anterior frontal voice area (aFVA) is most closely associated with the horizontal ascending ramus of the lateral fissure (half), which separates the *pars triangularis (area 45)* from *pars orbitalis (area 47/12)* in the anterior inferior frontal gyrus (Sprung-Much et al., 2020). It is located at the more anterior part of the inferior frontal sulcus (ifs), forming the dorsal border of area 45 (Frey et al., 2014). Last, it is also related to the anterior ascending ramus of the lateral fissure (aalf), which forms the caudal border of area 45, with area 44 located posteriorly (Sprung-Much et al., 2020). As such, aFVA is likely located in area 45 in the anterior part of Broca's speech region. The mFVA is often found close to the ifs, at the posterior part of the inferior frontal cortex (i.e., close to the *aalf* and diagonalis sulcus (ds)) relative to the aFVA. The main neighboring sulci of the mFVA form the boundaries of area 44 of the *pars opercularis* with *ifs* dorsally, *aalf* anteriorly, and inferior precentral sulcus *iprs* posteriorly, while the *ds* is known to be an axial sulcus within area 44 (Loh et al., 2017, 2020; Sprung-Much et al., 2018). However, in

some participants, the mFVA notably extends into the area 45 territory (i.e., near *half* and *ifs-anterior*). We propose that mFVA occupies *pars opercularis*, i.e., area 44. The pFVA is consistently located close to the *iprs,* which delimits area 44 anteriorly and ventral premotor area 6 posteriorly. It is sometimes found close to the *cs*, where the primary motor area 4 is found, and close to the *ifs*, which separates the middle frontal gyrus dorsally from the inferior frontal gyrus. Based on these neighboring sulci, the pFVA is in ventral premotor area 6 at the most caudal part of the inferior frontal cortex.

Limbic System The limbic system is crucial in processing emotional and social nuances embedded in vocal stimuli. It is an aggregation of brain structures generally located lateral to the thalamus, underneath the cerebral cortex, and above the brainstem. The anterior cingulate cortex (ACC) is situated in the frontal part of the cingulate gyrus, stretching from the corpus callosum's anterior segment to the cingulate sulcus's genu. The posterior cingulate cortex (PCC) is located posteriorly on the cingulate gyrus, extending from the cingulate sulcus's splenium to its isthmus. The amygdala is nestled within the medial temporal lobe, anterior to the hippocampus, and lateral to the thalamus. The insula, concealed by the lateral sulcus, lies between the temporal and frontal lobes, with its anterior part neighboring the frontal operculum and its posterior part adjoining the parietal operculum.

4. The fronto-temporal-limbic network of voice processing

Voice perception can be viewed from different angles. Some focus on how the brain functions when processing voices, looking at the cognitive processes and components of voice perception at a theoretical level (Belin et al., 2004). Others look at brain structure, studying the regions and pathways involved in recognizing voices (Belin et al., 2000; Staib & Frühholz, 2023). From a functional viewpoint, voices have mainly been studied as a multimodal neural network. This broader view comes from the analogous mechanisms by which the brain processes voice and face information, giving rise to the term “auditory face” (Campanella & Belin, 2007; Perrodin et al., 2015; Young et al., 2020). However, this term has its limits. Focusing only on the similarities between voice and face might miss some specific details about how voices are processed.

Studies over the past two decades have established via complementary neuroimaging techniques that the cerebral processing of voice information involves a set of temporal voice areas (TVAs) in secondary auditory cortical regions of the human (fMRI: Belin et al., 2000; von Kriegstein & Giraud, 2004; Pernet et al., 2015; EEG, MEG: Charest et al., 2009; Capilla et al., 2013; Barbero et al., 2021; Electrophysiology: Zhang et al., 2021; Rupp et al., 2022). The TVAs respond more strongly to sounds of voice — with or without speech (Pernet et al., 2015; Rupp et al., 2022; Trapeau et al., 2022)— and categorize voice apart from other sounds (Bodin et al., 2021).

Recent research indicates that while the TVAs are primarily associated with the auditory experience of voice through a general voice processing (Staib & Frühholz, 2021; Bestelmeyer & Mühl, 2022; Morillon et al., 2022; Staib & Frühholz, 2023), the frontal voice areas (FVAs) play a pivotal role in behaviorally significant voice processing tasks. These tasks include recognizing familiar voices, guided by focused attention and control. This underscores the importance of FVAs in models of voice perception (Aglieri et al., 2021; Roswandowitz et al., 2021; Bestelmeyer & Mühl, 2022). Another crucial facet of voice perception is emotional discernment. A growing body of recent studies underscores the integral role of the limbic system in voice processing, particularly in decoding emotional valence, reacting to voices, and modulating arousal and attention (Frühholz et al., 2019; Domínguez-Borràs et al., 2019; Giordano et al., 2021; Steiner et al., 2022). These additional areas —"the extended voice perception system"— are analogous to the extended network for face processing (Hesse & Tsao, 2020).

What is the functional role of each unit within the "voice patch" system in the human brain when processing vocal information? One leading theory about how we process sound suggests that the left side of the brain is better at handling quick changes in sounds, while the right side is better with slower changes (Zatorre & Belin, 2001; Flinker et al., 2019; Hamilton et al., 2019; Albouy et al., 2020; Morillon et al., 2022). This idea is supported by research showing speech processing happens mainly in the left hemisphere (Albouy et al., 2020) and recognizing who is speaking happens more in the right hemisphere (Mathias & von Kriegstein, 2014; Andics et al., 2010; von Kriegstein et al., 2003; Myers & Theodore, 2017; Hickok & Poeppel, 2007; Belin & Zatorre, 2003). This suggests that each side of the brain could be specialized for processing different aspects

of sound. Drawing from a range of studies —including behavioral, fMRI, M/EEG, and single-unit experiments (i.e., recording of spiking activity from a single neuron) — we put forth a synthesized functional model for voice. This model incorporates findings from previous studies suggesting that the right side of the brain predominantly processes voice identity information. In contrast, the left side focuses on deciphering voice semantic content and processing voice emotion bilaterally. Moreover, the model integrates the system responsible for processing vocal emotions. Furthermore, our model attempts to map these functions with the underlying neural networks involved in voice processing (Figure 1.1b).

How do voice patches connect within the brain? Blank et al. (2011) identified structural connectivity between voice-sensitive regions in the temporal lobe using diffusion-weighted imaging (DWI). More recently, Zhang et al. (2021) utilized ECoG electrode grids (i.e., placing electrodes directly on the exposed surface of the brain to record electrical activity, which provides a more direct and higher resolution measurement of brain activity than scalp EEG) and latency analyses and described dual voice processing pathways. In this proposal, information originates from the mTVA patches and bifurcates: one pathway leads from mTVA to aTVA, while the other extends from mTVA to pTVA. We hypothesize the existence of effective connectivity (EC, i.e., the causal link between different brain areas, that is, if the signal in one area influences the signal in another) among these patches, i.e., directional, causal neural interactions flow from mTVA to aTVA and pTVA. Aglieri et al. (2018) investigated functional connectivity (FC; i.e., the statistical association between neuronal activations in different regions of the brain, which helps understand how different parts of the brain communicate and work together during various tasks or at rest) within the voice perception network, defined by three frontal and three temporal regions of interest in each hemisphere, based on group voice-specific activation (the so-called temporal and frontal "voice patches"). They found that the TVAs and FVAs networks are functionally interconnected. Notably, in the right hemisphere, this connection proved significant for voice recognition performances.

What are the temporal dynamics of voice processing? Several studies employing MEG/EEG and intracranial recordings have illustrated that the mTVA exhibits selective responsiveness to voice commencing around 150-200 ms (Charest et al., 2009: 164 ms;

Capilla et al., 2013: 150 ms; Zhang et al., 2021: ~150 ms; Lowe et al. 2021: ~200 ms; Norman-Haignere et al., 2022: <200 ms; Rupp et al., 2022: ~150 ms). Additionally, several studies reported that the pTVA and aTVA manifest a longer response time compared to the mTVA, suggesting their position at a subsequent stage in the voice processing hierarchy (Schall et al., 2015: ~200 ms; Zhang et al., 2021: >200 ms; Norman-Haignere et al., 2022: >200 ms). Within the limbic system, the right amygdala and the right insula play a role in deciphering general emotional attributes. The right amygdala shows heightened activation during emotional state processing at a later stage, specifically post ~500 ms. In comparison, the right insula demonstrates increased activation post ~700 ms, marking a staggered temporal engagement in processing emotional states (Giordano et al., 2021).

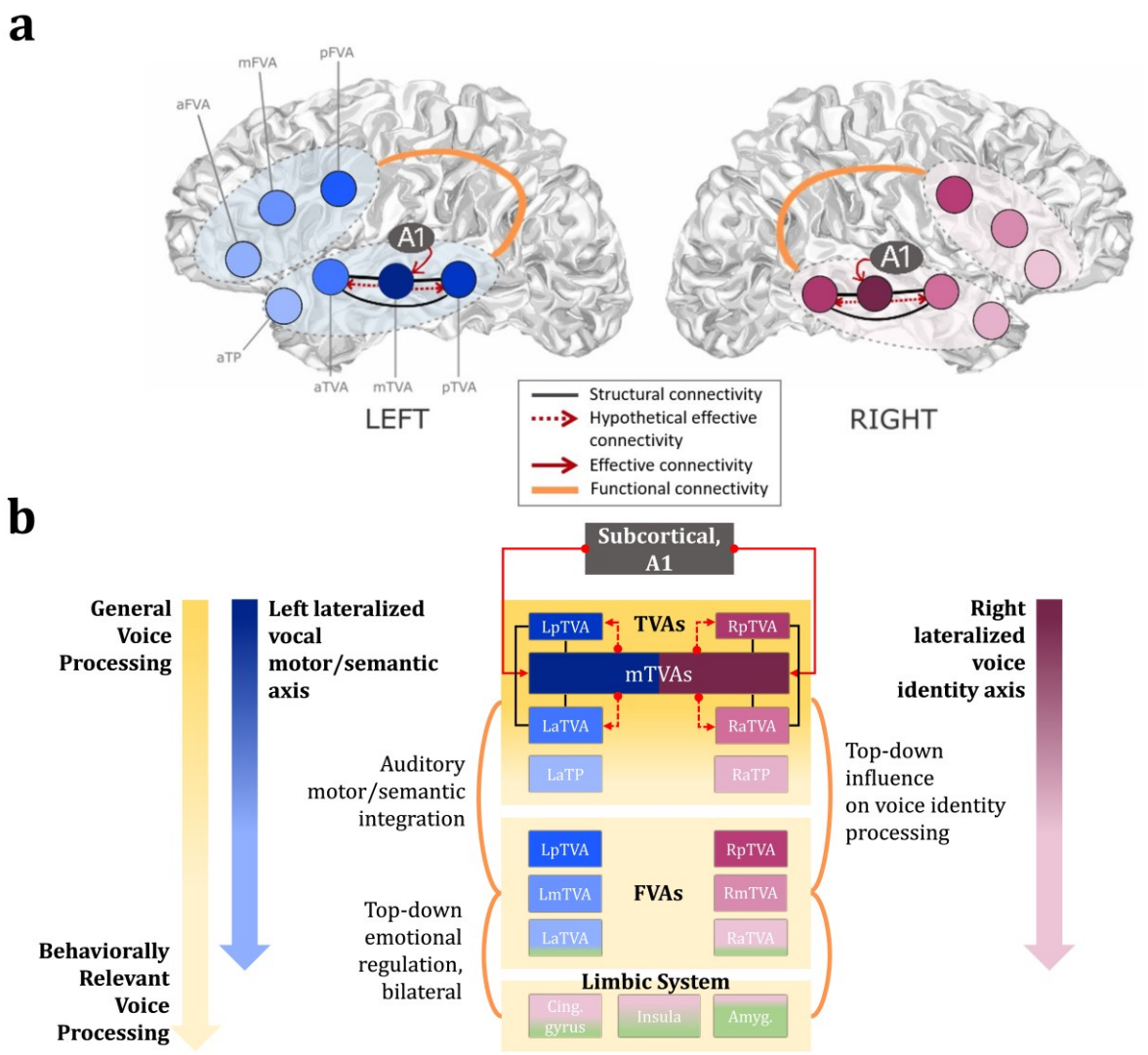


Figure 1.1: Schematic of the state-of-the-art knowledge of voice processing in humans. a, Brain regions responsive to voices in humans. Components shown in blue involve the vocal/motor processing axis in the left hemisphere. Components shown in pink involve the voice

identity processing axis in the right hemisphere. Lower-level analyses are indicated by more intense colors, determined approximately based on the literature. Black lines are used to indicate a structural connection. Fronto-temporal and fronto-limbic functional connections are indicated with orange lines. Plain and dotted red arrows indicate effective connectivity or hypothetical effective connectivity between two areas. Abbreviations: L/R, left/right hemisphere; A1, primary auditory cortex; a/m/pTVA, anterior/mid/posterior temporal voice area; aTP, anterior temporal pole; a/m/pFVA, anterior/mid/posterior frontal voice area. **b, Functional model of voice perception in humans.** The model shows components that involve unimodal responses to voices. Same colors as in **a**. Components in green involve the vocal emotion processing axis, bilaterally. Yellow indicates a general voice processing to behaviorally relevant voice processing gradient. Adapted from Belin et al. (2004), Maguinness et al. (2018), and Morillon et al. (2022).

5. General voice processing

A voice heard by a listener first undergoes general low-level auditory analyses, such as spectro-temporal filter analysis (Belin et al., 2000; Zatorre et al., 2002; Hickock et Poeppel, 2004; Hickok & Poeppel, 2007; Bodin et al., 2021; Rupp et al., 2022; Giordano et al., 2023), in subcortical areas and primary auditory cortex (A1). Then, a finer voice structural analysis (Staib & Frühholz, 2021) begins in the mid-temporal voice areas ("mTVA – Voice Structural Analysis"). Functionally, mTVA – bilaterally (Belin et al., 2000) – appears to perform a template matching to detect and match voices (or "norm-based coding"; Latinus et al., 2013) to an internal 'voice prototype' (Figure 1.2). In this perspective, neural responses do not mirror the stimulus directly; instead, they indicate its congruence with an internal template, a norm that could encapsulate the average of our personal voice experiences within our social context (Rupp et al., 2022).

Functionally, several recent studies have investigated the neuronal responses to voice stimuli in human nonprimary areas using intracranial recordings, either through ECoG electrode grids (Zhang et al., 2021) or sEEG recordings (Rupp et al., 2022). Their findings support the idea of a hierarchical organization of voice patches in the temporal lobe, where the information flow starts from the mTVA patches and moves in two directions: one from mTVA to the anterior TVA (aTVA) and the other one from mTVA to posterior TVA (pTVA).

Temporarily disrupting neuronal activity in the right mTVA using repetitive transcranial magnetic stimulation (rTMS; i.e., a non-invasive procedure that uses magnetic fields to stimulate nerve cells in the brain) impairs performance in voice detection tasks but not in broader auditory tasks (Bestelmeyer et al., 2011). This not only suggests a direct causal relationship with voice processing but also underscores RmTVA's higher hierarchical role in this process.

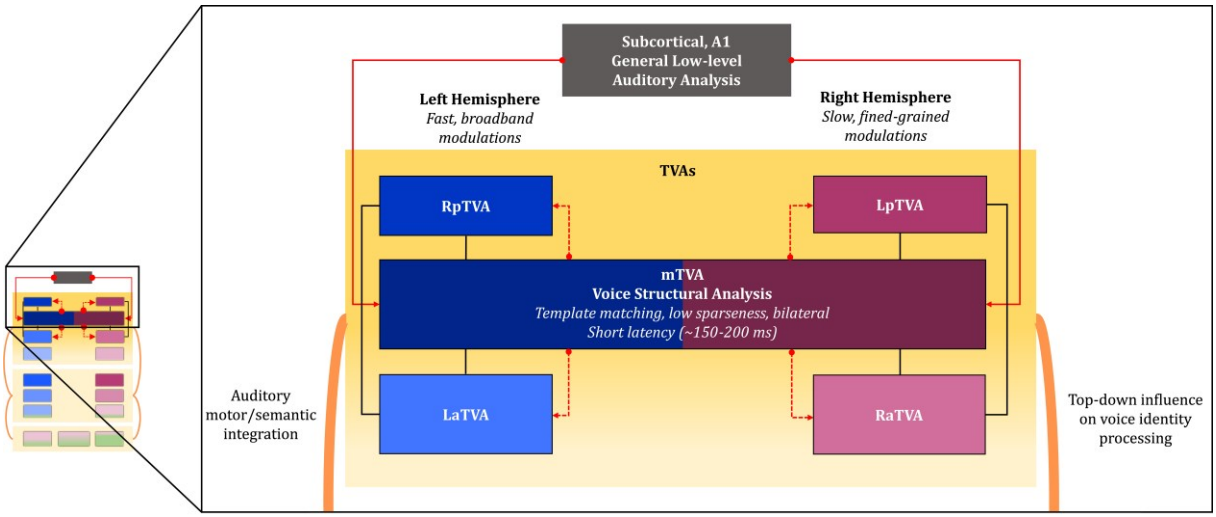


Figure 1.2: Schematic of the general voice processing in humans. A voice heard by a listener first undergoes general low-level auditory analyses in subcortical areas and primary auditory cortex (A1). Then, a finer voice structural analysis begins in the mid-temporal voice areas (mTVA – Voice Structural Analysis). Functionally, mTVA –bilaterally– appears to perform a template matching to detect and match voices to an internal 'voice prototype'. Recent findings support the idea of a hierarchical organization of voice patches in the temporal lobe, where the information flow starts from the mTVA patches and moves in two directions: one from mTVA to the anterior TVA (aTVA) and the other one from mTVA to posterior TVA (pTVA). Components shown in blue involve the vocal/motor processing axis in the left hemisphere. Components shown in pink involve the voice identity processing axis in the right hemisphere. Lower-level analyses are indicated by more intense colors, determined approximately based on the literature. Black lines are used to indicate a structural connection. Fronto-temporal functional connections are indicated with orange lines. Plain and dotted red arrows indicate effective connectivity or hypothetical effective connectivity between two areas. Yellow indicates a general voice processing to behaviorally relevant voice processing gradient. Abbreviations: L/R, left/right hemisphere; A1, primary auditory cortex; a/m/pTVA, anterior/mid/posterior temporal voice area.

6. Vocal motor/semantic processing axis

This axis is primarily left-lateralized, encompassing regions in both the temporal and frontal lobes, which play vital roles in vocal and semantic processing. The left TVAs might first realize a general vocal motor/semantic processing, while the left FVAs might integrate this information via frontotemporal connections (Figure 1.3).

Left TVAs: General Vocal Motor/Semantic Processing Tsapkini et al. (2011) emphasized the potential role of the anterior temporal pole (aTP) as a semantic hub for semantic knowledge. Based on assessments of acute strokes and infarct volumes, they found that both the right and left aTP are involved in processing and understanding meanings and concepts, such as words. Functionally, the left anterior temporal voice area (LaTVA) might be involved in semantic processing (Patterson et al., 2007; Perrodin et al., 2015; Zhang et al., 2021) and correlates with motor areas (Aglieri et al., 2018). It also contributes to formant tracking (Latinus et al., 2013). Cope et al. (2020) proposed that LaTVA is crucial for the efficient lateralized processing of spoken word identity. The posterior temporal voice area (pTVA) might play a crucial role in auditory-motor integration. Functionally, the pTVA encodes phonetic features (von Kriegstein et al., 2010; Mesgarani et al., 2014; Zhang et al., 2021) and exhibits high latency and sparseness (Zhang et al., 2021). The LpTVA, often termed "motor speech", correlates with motor regions during speech processing (von Kriegstein et al., 2010; Zhang et al., 2021).

Left FVAs: Auditory Motor/Semantic Integration In the left hemisphere, the aFVA connects to (1) Higher-order processed semantic and multimodal inputs from the anterior and middle parts of the temporal lobe through the extreme capsule fasciculus (Frey et al., 2008; Petrides & Pandya, 2009), forming the ventral speech pathway (Hickok & Poeppel, 2007); (2) Auditory inputs processed from the posterior temporal cortex via the arcuate fasciculus (Frey et al., 2014); (3) Speech output areas such as area 44 and the ventral premotor areas positioned more posteriorly. This connectivity implies that the LaFVA might retrieve and integrate higher-order semantic and auditory aspects of voice information. It then might guide speech motor actions via the posterior inferior frontal cortex (Loh et al., 2020). Non-human primates likely share this functionality, given the consistent connectivity of area 45 across both species. The mFVA might then be part of the dorsal stream of speech perception (Erickson et al., 2017) and could be connected

with social cognition processes (Hamzei et al., 2016). In humans and monkeys, this region communicates with the posterior temporal cortex via the arcuate fasciculus and the supramarginal gyrus through the superior longitudinal fasciculus III. In the left hemisphere, this network comprises the dorsal speech pathway (Hickok & Poeppel, 2007), which is implicated in the phonological processing of speech. Area 44, where LmFVA resides, associates closely with the speech motor output region in the ventral premotor cortex (Petrides et al., 2014). Thus, the LmFVA might process phonological aspects of voice and modulate control over speech/vocal motor production via the ventral premotor area. This functionality is backed by studies showing area 44's involvement in selecting orofacial and vocal motor (Loh et al., 2020). Non-human primates exhibit similar attributes, with area 44 involved in auditory-driven vocal motor control (Aboitiz, 2018; Hage & Nieder, 2013). The pFVA, known for high-level motor speech and voice identity representation, might house speech-motor representations (Conant et al., 2014). Guenther et al. (2017) suggested that the left ventral premotor area might offer a top-down perception of speech by formulating predictive models of speech motor plans. These models then juxtapose with perceived auditory-vocal inputs in the temporal cortex. Given its anatomical ties to the posterior parietal and temporal cortices and its position in the dorsal speech pathway in the left hemisphere described by Hickok and Poeppel (2007), the LpFVA's role in voice processing probably involves a top-down influence on speech-vocal perception.

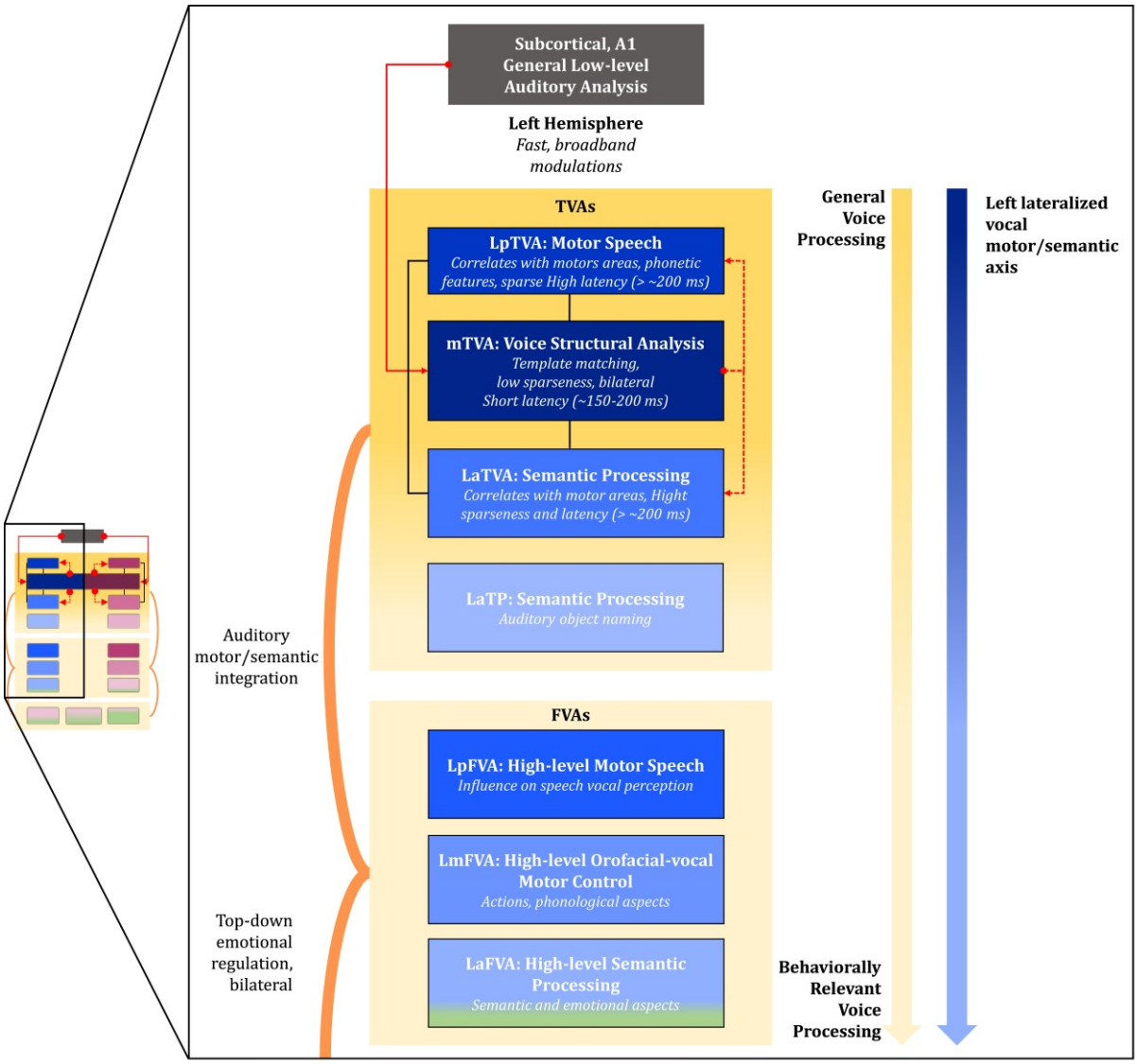


Figure 1.3: Schematic of the vocal motor/semantic processing axis. Components shown in blue involve the vocal/motor processing axis in the left hemisphere. Components in green involve the vocal emotion processing axis, bilaterally. Lower-level analyses are indicated by more intense colors, determined approximately based on the literature. Black lines are used to indicate a structural connection. Fronto-temporal and fronto-limbic functional connections are indicated with orange lines. Plain and dotted red arrows indicate effective connectivity or hypothetical effective connectivity between two areas. Yellow indicates a general voice processing to behaviorally relevant voice processing gradient. Abbreviations: L, left hemisphere; A1, primary auditory cortex; a/m/pTVA, anterior/mid/posterior temporal voice area. aTP, anterior temporal pole; a/m/pFVA, anterior/mid/posterior frontal voice area.

7. Voice identity processing axis

How is voice identity encoded in the brain? Again, this axis is primarily right-lateralized, with the temporal and frontal lobes as the main regions. The right TVAs might

first realize general voice identity processing, while the left FVAs might integrate this information via frontotemporal connections into higher-level representations, as well as influence the temporal representations during active voice identity recognition (Figure 1.3).

Right TVAs: General Voice Identity Processing Several studies have explored the involvement of the anterior temporal pole (aTP) in voice identity processing. Belin and Zatorre (2003) demonstrated adaptation to a speaker's voice in the right hemisphere. Antics et al. (2010) suggested bilateral identity processing, whereas Latinus et al. (2011) discussed learning-induced changes in the cerebral processing of voice identity in the right hemisphere. Additionally, Luzzi et al. (2018) observed selective associative phonagnosia (a condition where an individual has difficulty recognizing familiar voices despite having normal hearing and speech perception abilities) following a right anterior temporal stroke, emphasizing familiar identity representation. Drawing on the studies above, it can be inferred that the right hemisphere of the anterior temporal pole (aTP) could correspond to the supramodal person identification stage as proposed in the functional model of Belin et al. (2004). Zhang et al. (2021) noted that the aTVA possesses high latency and sparse activations compared to mTVA, indicating the information transfer from mTVA. Functionally, the left aTVA is involved in semantic processing (Patterson et al., 2007; Perrodin et al., 2015; Zhang et al., 2021) and correlates with motor areas (Aglieri et al., 2018). It also contributes to formant tracking (Latinus et al., 2013). The right aTVA might primarily represent human voice identity (Maguinness et al., 2018; Zhang et al., 2021) and is associated with identity adaptivity (Kriegstein & Giraud, 2004), timbre discernment (Pernet & Belin, 2012; Allen et al., 2017), and f0 tracking (Schuller, 2013). Sometimes referred to as a “person identity node”, the right pTVA has strong connections with the anterior facial voice areas (aFVAs). This connectivity might assist in determining who is speaking (Aglieri et al., 2018).

Right FVAs: Top-down Influence on Voice Identity Processing The RaFVA has been associated with the processing of vocal attractiveness (Bestelmeyer et al., 2012) and the representation of voice gender (Charest et al., 2013). RmFVA has been associated with the processing of vocal attractiveness (Bestelmeyer et al., 2012). In the right hemisphere, the pFVA is associated with perceiving voice identity despite acoustic variability, especially

after familiarization with previously unfamiliar voices (Latinus et al., 2011; Antics et al., 2013). This supports the idea of top-down influences in voice identity processing.

Limbic System The anterior insulae were involved in the higher-level representation of identity. This supports the hypothesis that in tandem with the medial superior frontal region, it supports person identity recognition (Bestelmeyer & Mühl, 2022). The left cingulate gyrus has been found to be sensitive to changes in perceived voice identity, suggesting its role in storing and retrieving familiar voices (Latinus et al., 2011). Both the cingulate gyrus's left and right anterior portions demonstrate activations in response to familiar voices (von Kriegstein & Giraud, 2004). The anterior and posterior left areas are associated with voice gender processing, especially in perceived ambiguous voices (Charest et al., 2013). Blank et al. (2014) found that the left anterior region is involved in recognizing well-known voices, while the right anterior region is associated with recognizing familiar voices. The bilateral insula is associated with voice gender recognition tasks (Charest et al., 2013). The left amygdala is involved in processing voice identity (Andics et al., 2010). The anterior insulae were involved in higher-level representation of identity, supporting the hypothesis that it supports person identity recognition in tandem with the medial superior frontal region (Bestelmeyer & Mühl, 2022).

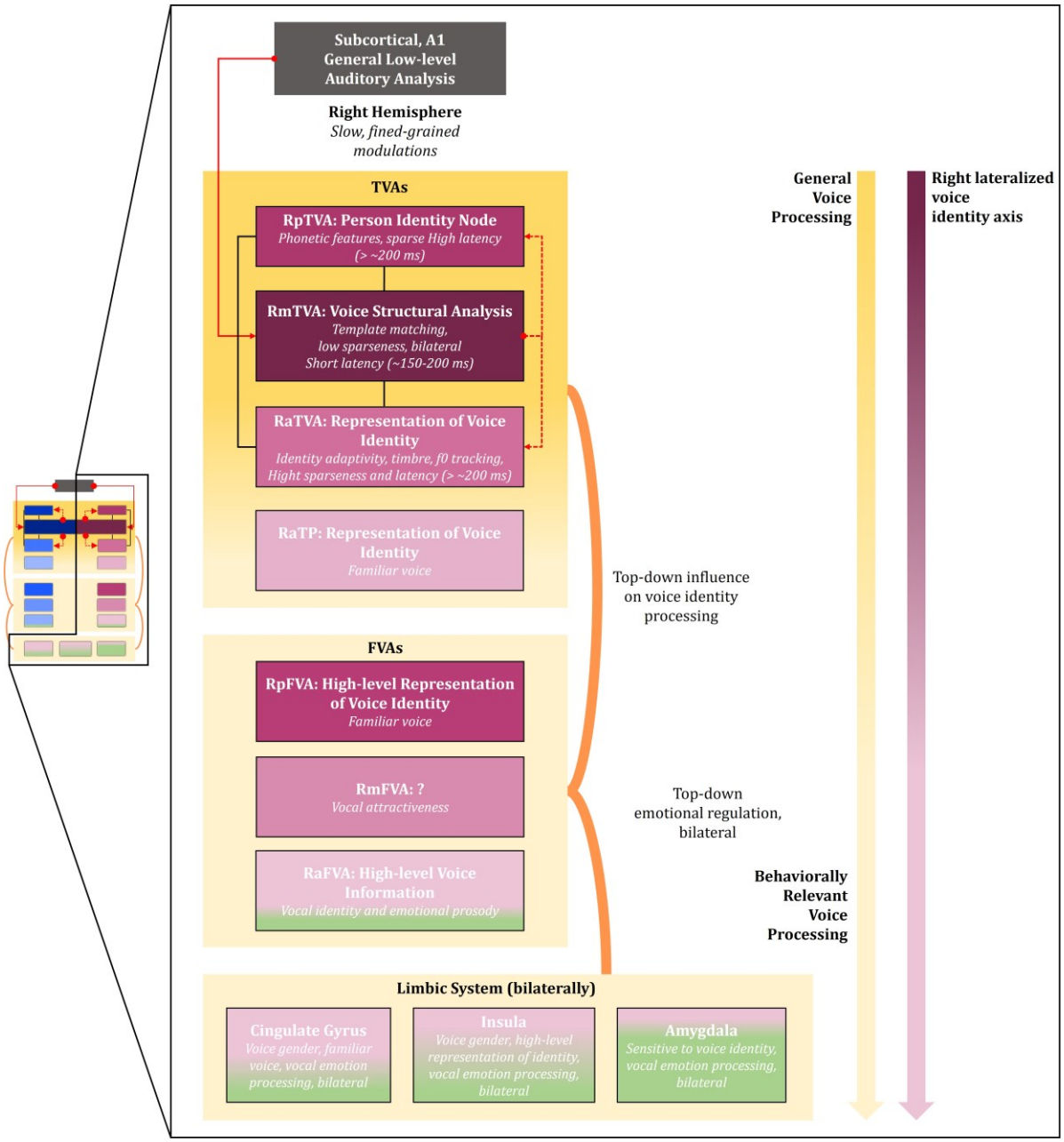


Figure 1.4: Schematic of the voice identity processing axis. Components shown in pink involve the voice identity processing axis in the right hemisphere. Components in green involve the vocal emotion processing axis, bilaterally. Lower-level analyses are indicated by more intense colors, determined approximately based on the literature. Black lines are used to indicate a structural connection. Fronto-temporal and fronto-limbic functional connections are indicated with orange lines. Plain and dotted red arrows indicate effective connectivity or hypothetical effective connectivity between two areas. Yellow indicates a general voice processing to behaviorally relevant voice processing gradient. The question mark indicates an undefined functional role for the corresponding region. Abbreviations: L, left hemisphere; A1, primary auditory cortex; a/m/pTVA, anterior/mid/posterior temporal voice area. aTP, anterior temporal pole; a/m/pFVA, anterior/mid/posterior frontal voice area.

8. Vocal emotion processing axis

aFVA: High-level Semantic and Voice Information Processing The RaFVA has been associated with detecting emotional tones in voices (Frühholz et al., 2012). Furthermore, this area is involved with representing emotional states (Giordano et al., 2021).

Limbic System: Vocal Emotion Processing The limbic system drives, provides, and identifies emotional and attitudinal elements in the voice (Robinson, 1976). Specifically, the amygdala plays a crucial role in processing vocal emotional information in humans (Frühholz et al., 2015; Frühholz & Grandjean, 2013). The left and right anterior cingulate gyri participate in vocal emotion processing (Ceravolo et al., 2021). In the insula, the correct region is involved in processing general emotional attributes and demonstrates increased activation during emotional state processing after about 700 ms (Giordano et al., 2021). Additionally, both sides of the insula are engaged in vocal emotion processing (Ceravolo et al., 2021). In voice processing, the amygdala has mainly been associated with the processing of emotional information in the voice in humans (Frühholz et al., 2015; Frühholz & Grandjean, 2013) or acoustic cues like roughness (Arnal et al., 2015). The right amygdala is involved in representing dimensions of emotion (more than specific categories) and processing emotional states after ~500 ms (Giordano et al., 2021). The left amygdala is involved in emotional cues in speech (Steiner et al., 2022; Frühholz, Hofstetter, et al., 2015; Anderson & Phelps, 2001). Bilaterally, the amygdala processes emotional voices, particularly distinguishing between fearful and neutral tones (Domínguez-Borràs et al., 2019; Frühholz et Grandjean, 2013).

However, it should be noted that some studies reported responses to neutral stimuli (i.e., no emotional content, no task related to identity recognition) in both the left and right amygdala regions (Pernet al., 2015; Aglieri et al., 2018).

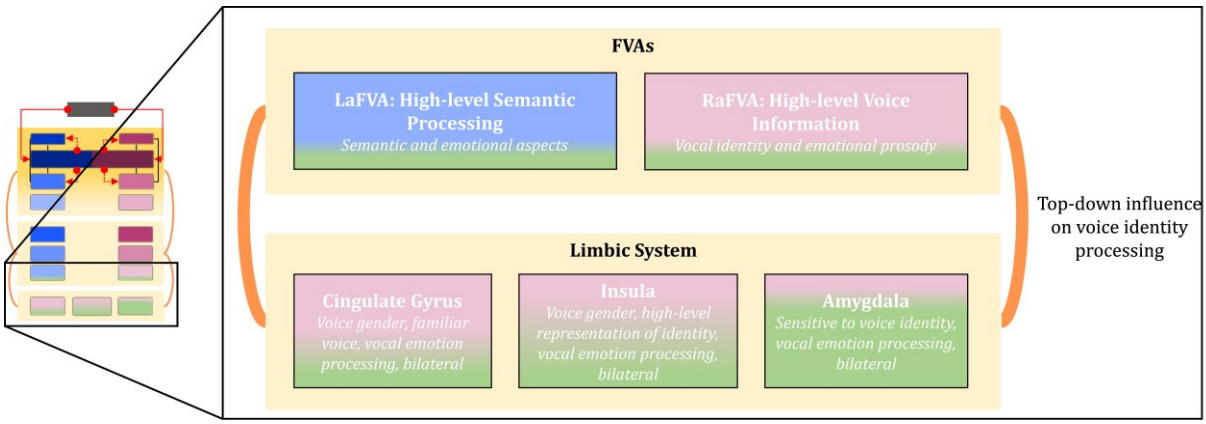


Figure 1.5: Schematic of the vocal emotion processing axis. Components shown in blue involve the vocal/motor processing axis in the left hemisphere. Components shown in pink involve the voice identity processing axis in the right hemisphere. Components in green involve the vocal emotion processing axis, bilaterally. Lower-level analyses are indicated by more intense colors, determined approximately based on the literature. Fronto-limbic functional connections are indicated with orange lines. Light yellow background color indicates behaviorally relevant voice processing. Abbreviations: L/R, left/right hemisphere; A1, primary auditory cortex; a/m/pFVA, anterior/mid/posterior frontal voice area.

9. Voice patch system across primate brains

Is there a shared voice coding principle across primate species? Voices play a crucial role in the social dynamics of many species. For a complete understanding of various social behaviors, it is essential to scrutinize vocal behavior, especially in mammals. Non-human primates, our closest evolutionary relatives, show comparable patterns in processing vocal information, both at the behavioral and neurological levels. By studying various primate species, we can investigate the origins of vocal perception. This enables us to trace changes since our last common ancestor and explore vocal perception mechanisms over time. Belin et al. (2018) suggested that the primate brain processes vocal information through a "voice patch system". We gather insights from voice processing in other primates to identify shared functional patterns or homologies, as outlined in Figure 1.6.

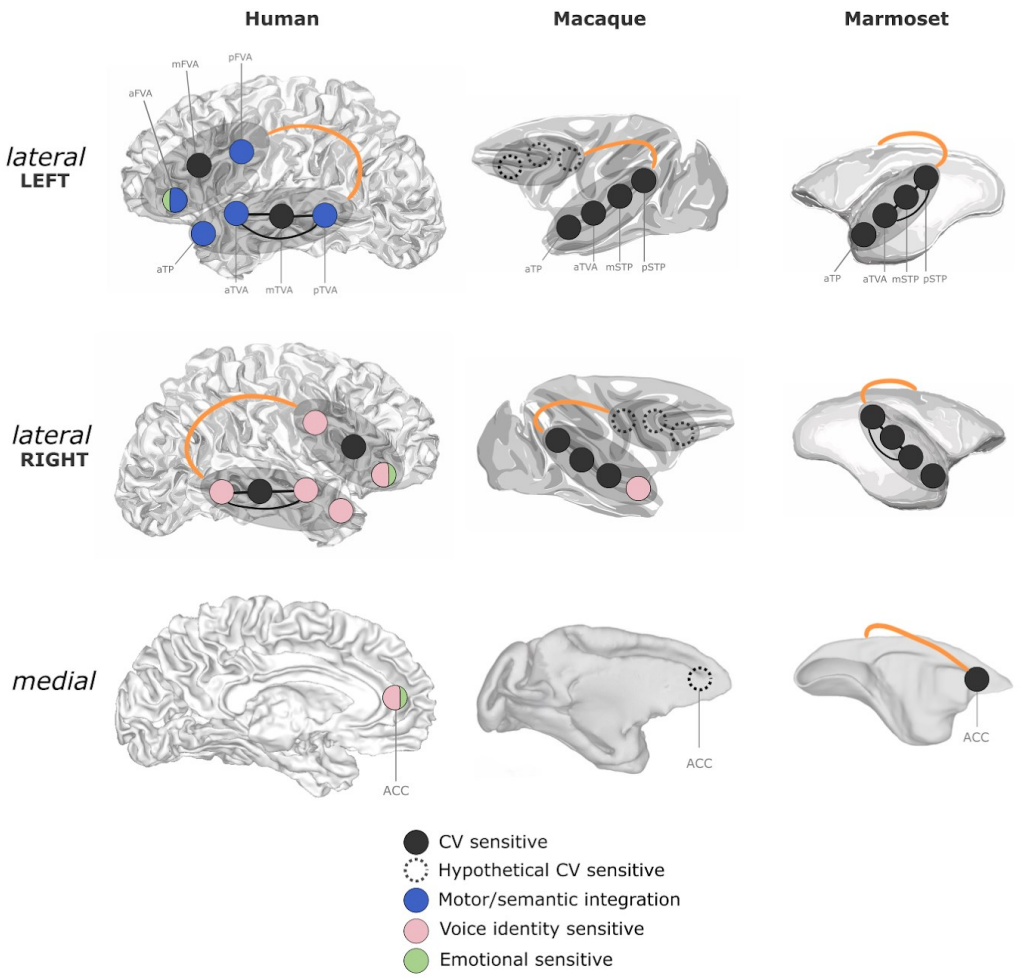


Figure 1.6: Brain regions responsive to voices in humans and non-human primates. Components shown in black involve conspecific vocalizations (CV) sensitivity. Components shown in blue involve the vocal/motor processing axis in the left hemisphere. Components shown in pink involve the voice identity processing axis in the right hemisphere. Components shown in dotted lines are hypothetical. Abbreviations: L/R, left/right hemisphere; A1, primary auditory cortex; a/m/pTVA, anterior/mid/posterior temporal voice area; aTP, anterior temporal pole; a/m/pFVA, anterior/mid/posterior frontal voice area; ACC, anterior cingulate cortex. Black and orange lines are used to indicate a structural or a functional connection, respectively.

Functional Responses to Conspecific Vocalizations (CVs) The initial step in identifying homologies in the voice patch system across primate brains is determining areas sensitive to CVs. The temporal lobe is central to vocalization processing in both macaques and marmosets. Early fMRI studies determined that the anterior temporal pole (aTP), along with patches resembling pTVA and mTVA in macaques, exhibit a clear preference for CVs (Petkov et al., 2008; Joly et al., 2012; Ortiz-Rios et al., 2015). The voice patch in the anterior temporal lobe (aTVA) has been consistently observed bilaterally and

demonstrates pronounced voice selectivity (Petkov et al., 2008; Perrodin et al., 2011; Ortiz-Rios et al., 2015; Bodin et al., 2021). The only two groups that investigated the activity of single neurons in macaque TVAs confirmed the leading role of aTVA in the processing of CVs (Perrodin et al., 2011; Giamundo et al., *submitted*). Using an fMRI-guided electrophysiological technique, these studies reported the existence of neurons selective to CVs that categorize CVs apart from other sounds. However, Giamundo et al. (*submitted*) also observed a population of aTVA neurons exhibiting selectivity towards human voices (sounds of paramount relevance in laboratory macaques' auditory environment), suggesting that aTVA neuronal activity can also represent vocalizations from other species with which primates have developed expertise. Similarly, marmosets show bilateral responsiveness to vocalizations in aTVA, mTVA, and pTVA, which might be analogous to human counterparts (Sadagopan et al., 2015; Jovanic et al., 2022; Jafari et al., 2023). An “extended voice system” also appears present in non-human primates. In macaques, the prefrontal cortex is notably sensitive to their vocalizations (Cohen et al., 2009; Romanski et al., 2005). This is reflected in marmosets where vocalizations activate the primary motor cortex, somatosensory cortex, and various prefrontal areas such as 8aV, 6DR, and 6M (Jafari et al., 2023; Jovanovic et al., 2022; Miller et al., 2015). However, marmosets do not exhibit strong selectivity in the human language network's inferior frontal cortex, highlighting potential evolutionary divergences in primate vocal processing (Jafari et al., 2023).

Functional Roles Several studies leverage intracranial recordings in macaques to explore the hierarchical organization of vocalization processing. Fukushima et al. (2014) described a progression in the neural coding of vocalizations along the ventral auditory pathway. In this pathway, rostral areas require the amalgamation of spectral and temporal features. Similarly, Kikuchi et al. (2010) identified a hierarchically organized auditory processing stream in the supratemporal plane (STP) that spans from the primary auditory area to the temporal pole, showcasing an increased stimulus specificity. It suggests that, like humans, macaques process CVs hierarchically along the rostral direction of the auditory cortex. Jafari et al. (2023) observed the presence of a voice processing network in marmosets, particularly within the rostral sections of the anterior cingulate cortex (ACC). In humans, this region has been associated with various voice-processing tasks, including voice learning (Latinus et al., 2011), recognizing familiar

voices (von Kriegstein & Giraud, 2004; Blank et al., 2014), discerning voice gender (Charest et al., 2013), and interpreting vocal emotions (Ceravolo et al., 2021). Given that the auditory stimuli in this study were familiar conspecific vocalizations, there might be a connection to the same processes observed in humans. This notion hints at an evolutionarily conserved processing mechanism, warranting further investigation in both macaques and marmosets. In the same study by Jafari et al. (2023), it was observed that the marmosets' pTVA had robust connectivity with the motor and somatosensory cortices, akin to humans (Frey et al., 2014). This similarity hints that the pTVA in marmosets might be involved in auditory-motor integration, as suggested in humans (von Kriegstein et al., 2010).

Functional Homologies Understanding the functional roles of voice units in the primate brain is more limited than in humans. Bodin, Trapeau, et al. (2021) employed comparative fMRI, highlighting that humans and macaques have bilateral voice areas in the anterior temporal lobe (aTVA). These areas show a preference for conspecific vocalizations and demonstrate a representational geometry that distinctly categorizes them from other sounds. This species-specific but homologous manner of categorization confirms earlier findings regarding speaker adaptation in the right aTP (Petkov et al., 2008).

Neural Connectivity Connectivity patterns in the primate brain support the idea of homologies. The frontotemporal network plays a role in vocal communication (Balezeau et al., 2020; Rocchi et al., 2021). It has two main pathways: the postero-dorsal and antero-ventral, comparable to the dorsal and ventral streams in the visual system (Rauschecker & Scott, 2009). The posterodorsal pathway, connecting the caudal belt of the auditory cortex to the dorsolateral PFC (dlPFC), is associated with the spatial processing of auditory signals. The anteroventral pathway, which connects the anterior belt of the auditory cortex to the ventrolateral PFC (vlPFC), is implicated in encoding different call types (Rauschecker & Scott, 2009). While evidence for these pathways exists in rhesus macaques, research on marmosets is limited. Some studies suggest marmosets may have a similar system (Grijseels et al., 2023), but more research is needed. In their 2023 study, Jafari et al. employed functional connectivity (FC) and tracer-based cellular connectivity (a technique used to study the pathways and connections between neurons) to investigate

marmosets' functional and structural links. They discovered that the three temporal voice patches were functionally interconnected and connected to the anterior cingulate cortex, especially area 32. The mTVA demonstrated functional connectivity with frontal areas 8aD, 8aV, and 47L, while pTVA exhibited robust connections with the motor and somatosensory cortices, a pattern also observed in humans. Their tracer-based findings further revealed that the anterior cingulate's area 32 maintained strong structural and functional ties with other cortical and subcortical vocalization-processing regions in marmosets.

10. Voice recognition mechanisms

In 2008, Tsao and Livingstone proposed a face recognition model involving three main computational steps. Firstly, there is a detection phase where we recognize something as a face. Then, we analyze the face to pinpoint its unique features. Lastly, using these features, we categorize the face based on identity, gender, age, race, and expression. In this model, detecting a face and identifying its specifics are distinct. To identify a face, we focus on what makes it different from others, even though all faces have general similarities. However, we are primarily concerned with the shared features of all faces for detection. This means a system efficient at detection might not excel in detailed identification, and vice versa.

What are the voice recognition mechanisms? We hypothesize that voice recognition operates on a similar principle: first, there is the detection of a voice, followed by an analysis of its distinctive characteristics, and then its categorization, e.g., based on the speaker's identity or emotional state. This is a complex task due to varying factors like pitch, tone, volume, and background noise. As with faces, we can segment voice recognition into these three main computational phases.

Detection At the most basic level, detection involves recognizing the presence of a voice by identifying shared auditory features. Common characteristics of voices, such as pitch and timbre, are crucial here. The challenges of detecting a voice and identifying its unique traits are separate in voice recognition. We focus on the differences in individual voice identification, even as all voices have commonalities. Conversely, detection is about pinpointing shared characteristics. If a system is good at simple detection, it might

struggle with detailed identification and vice versa (Tsao & Livingstone, 2008). Moreover, detection serves as a filter, activating detailed voice recognition processes only when a sound qualifies as a voice. This domain-specific gating may be one reason for the brain's anatomical segregation of voice processing. Another advantage of this detection step is that it distinguishes the voice from any background noise, aligning it for further analysis. Several voice recognition algorithms depend on this initial separation, especially when faced with irregular backgrounds (Singh et al., 2018). Computer algorithms of voice activity detection (VAD) mainly relied on cepstral-based algorithms—such as Mel-frequency cepstral coefficients (MFCCs)—as they exhibit a high degree of independence to levels of background noise (Haigh & Mason, 1993; Wang et al., 2011). Recently, end-to-end deep neural networks (DNNs) have been utilized to map acoustic inputs directly to predefined semantic categories—such as human voice, music, and natural sounds—by leveraging large datasets, ranging from thousands to hundreds of thousands of hours annotated with human labels (Gemmeke et al., 2017; Hershey et al., 2017). This advancement renders voice detection more realistic in natural and noisy environments. We identified above (see *General Voice Processing*) that in the brain, the bilateral mTVAs in the human brain could act as a template matching to detect and match voices (Latinus et al., 2013). Their causal link with voice detection has been established by transiently interfering with neuronal activity in the right TVAm via transcranial magnetic stimulation (TMS) interferes with performance at a voice detection task but not at a more general auditory task (Bestelmeyer et al., 2011).

Measurement Upon detection of a voice, it requires measurement in a way that enables accurate, efficient identification. The measurement process should not be so coarse that it misses the subtle features differentiating one voice from another. Alternatively, it should yield a set of values that can be efficiently juxtaposed with stored templates for identification purposes. A zero-sum game exists between measurement and categorization: the more streamlined the measurement, the simpler the classification; conversely, less efficient measurement renders the classification process more demanding (Tsao & Livingstone, 2008). Deep Neural Networks (DNNs)-based classifiers (LeCun et al., 2015) exemplify this: the input undergoes a long hierarchical series of highly nonlinear transformations (measurements), while the final classification layer is often a simple linear transformation (categorization). During the vocal processing, once a voice

has been detected and undergone a preliminary general preprocessing by mTVA, the information bifurcates: one pathway leads from mTVA to aTVA, while the other extends from mTVA to pTVA (Schall et al., 2015; Zhang et al., 2021). Depending on the domain of required expertise, different parts of the information will be processed by different brain regions. The general processing of speaker identity begins in the right pTVA and aTVA (see voice identity axis), the general semantic processing begins in the left pTVA and aTVA (see motor/semantic axis), and the vocal emotion processing is mainly in the limbic system, bilaterally. Simultaneously, the TVAs of each respective axis guide the general processing with top-down influence, especially in behaviorally relevant voice processing (e.g., speaker recognition). As the information is processed along this hierarchical stream, the representations associated with the different axes— the motor/semantic representations, the speaker identity representations, and the emotional representations—become of increasingly higher orders and are stored at different locations. The right pTVA might be responsible for establishing identity patterns, while the right aTVA has been suggested to encode higher-order representations compared to the other temporal VA. Indeed, Luzzi et al. (2018) observed selective associative phonagnosia following a right anterior temporal stroke, with the correct part potentially corresponding to the supramodal person identification stage as proposed in Belin et al. (2004) functional model. In contrast, the left encodes high-level semantic information, such as auditory object naming.

Categorization Separating the measurement process from the classification process gives a computational system maximum flexibility because different categorizations (e.g., speech, speaker's identity, or emotional state) can all operate from the exact representation. Based on the evidence we gathered in our synthesized model, the categorization step might be performed in the "extended voice perception system" (Antics et al., 2010; Latinus et al., 2011; Charest et al., 2013; Antics et al., 2013; Blank et al., 2014; Frühholz, Hofstetter, et al., 2015; Zäske et al., 2017; Luzzi et al., 2018; Ceravolo et al., 2021; Aglieri et al., 2021; Giordano et al., 2021; Bestelmeyer & Mühl, 2022; Steiner et al., 2022).

11. Using deep networks to probe representations in voice patches

Which computational models align with the VA representations? In the previous section (Voice recognition mechanisms), we identified deep neural networks (DNNs). A prevailing notion is that the brain learns largely unsupervised, constructing representations that elucidate the structure implicit in the raw sensory input (Lillicrap et al., 2020). Autoencoders are one example of learning such kinds of representations for voice—after that, they are named the 'voice latent space' (VLS). It learns to compress voice stimuli with high dimensionality into a lower-dimensional space that allows reconstruction of the original voice stimuli via an inverse transformation learned by the second part of the network called the decoder. Once such a lower-dimensional representation of voice is learned, we could linearly map it with the brain responses to voice stimuli.

12. Conclusion

This chapter reviewed both older and recent literature on voice processing in human and non-human primates to determine the potential role of each voice-sensitive area. We proposed a synthesized voice processing model based on brain studies in primates that outlines a pathway with three stages: detection, measurement, and categorization for voice recognition. The model tentatively underscores the roles of the fronto-temporal-limbic network and the hemispheric specialization, where the right predominantly handles voice identity, the left manages semantic deciphering, and the limbic system, the vocal emotion, bilaterally.

However, several key questions remain to be elucidated. ***How is voice identity encoded in the brain?*** Although we identified a potential voice identity processing axis with a candidate functional role and a tentative degree of abstractness of the representations (Figure 1.4), the exact computations performed in these representations are still unknown, particularly regarding voice identity information. We propose to explore this question by mapping the brain responses to voice stimuli recorded in the VA with the representations learned by DNNs, as demonstrated in Chapter 3. Besides, other important information is still missing, e.g., ***what are the structural/functional***

connections within the frontal areas? Potential future research is discussed in the Discussion (Section *Evolutionary origins of voice perception*).

Is there a shared voice coding principle across primate species? To extend my proposal to use DNNs, and in general AI, as a model to probe the representations in the vocal brain (Chapter 3), one would need sufficient vocal samples to train this kind of model. In the next chapter (Chapter 2), I show how to use AI as a tool to build a large dataset of non-human primate vocalizations in a semi-supervised fashion.

Chapter 2

Towards studying the evolution of vocal communication systems with deep learning

In this chapter, I propose an end-to-end pipeline for processing vocalizations from raw recordings of marmoset monkeys. This includes detection, segmentation, and labeling. This dataset will be the first milestone in future studies to train efficient computational models, such as DNNs, to learn high-level representations of monkey vocalizations.

1. Abstract

As our closest relatives, non-human primates use a wide range of complex vocal signals for communication within their species. Previous research on marmoset (*Callithrix jacchus*) vocalizations has been limited by recording setups with low sampling rates and insufficient labeling for advanced analyses using Deep Neural Networks (DNNs). Here, we provide a database of common marmoset vocalizations, continuously recorded with a sampling rate of 96 kHz from a stabulation room housing ~20 marmosets in three cages simultaneously. The dataset comprises over 800,000 files, amounting to 253 hours of data collected over 40 months. Each recording lasts a few seconds and captures the marmosets' social vocalizations, encompassing their entire known vocal repertoire during the experimental period. Around 215,000 calls are annotated with the vocalization type. The dataset presented here contributes to our understanding of voice phylogeny by providing a more detailed characterization of the acoustic properties of the marmoset vocal repertoire. These data hold the potential for shedding light on the origins of syntax, semantics, and the evolution of vocal communication systems. Furthermore, we offer a trained classifier to assist future investigations.

2. Introduction

Non-human primates, the closest evolutionary relatives to humans, exhibit various complex behaviors, including the extensive use of acoustically diverse vocal signals for communication within conspecifics. By conducting comparative research on non-human primates, valuable insights can be gained into the evolutionary development of speech and language. For example, studying their vocal communication can provide clues about the origins of syntax and semantics. Although non-human vocalizations can be complex for humans to decipher, large acoustic datasets may make it possible to identify essential nuances critical to animal communication but imperceptible to the human ear. There has been considerable interest recently in the common marmoset (*Callithrix jacchus*) as a neuroscientific model organism (Miller et al., 2016), and many attempts have been made to study and characterize its vocal repertoire (Epple, 1968; Pistorio et al., 2006; Bezerra et al., 2008; Agamaite et al., 2015; Zhang et al., 2018; Zhao et al., 2019). However, among the past and recent literature, the audio recording setups did not allow recording above a sampling rate of 48 kHz which would allow the entire frequency range of marmoset vocalizations, corresponding to their hearing range from 125 Hz to 36 kHz (Osmanski & Wang, 2011), to be recorded.

Furthermore, the existing datasets do not provide a sufficient number of labeled vocalizations to leverage advanced analytical methods. Fine-grained statistical analyses, such as those based on Deep Learning for decoding animal communication, require substantial data (Rutz et al., 2023).

Here, we present an extensive collection of vocalizations of marmosets. We have acquired and segmented over 800,000 vocalizations with a sampling rate of 96 kHz from a stabulation room containing 3 cages (~20 marmosets) over three years. Marmosets are capable of producing a diverse array of vocalizations, including trills, phees, twitters, tsiks, seeps, and infant cries, even when kept in captivity (Bezerra & Souto, 2008; Epple, 1968; Remington et al., 2012; Rylands, 1993). Identifying the pertinent voiced segments within a recorded audio track is frequently the primary hurdle in audio data analysis. To address this, we made use of signal processing and deep learning tools to segment automatically and cluster vocalizations based on the methods described in the recent computational neuroethology literature (Sainburg et al., 2020; Sainburg & Gentner, 2021; Best et al., 2023). The comprehensive dataset we present has the potential to improve our comprehension of voice phylogeny by better characterizing the acoustical properties of the marmoset vocal repertoire, e.g., by comparing the sequential organization of acoustic elements across species (Sainburg et al., 2019).

3. Results

3.1. Data records

The data consist of:

1. 869,556 recorded audio files (253 hours; FLAC format, sampling rate: 96 kHz, depth: 32 bit).
2. One annotation file: Annotations.tsv, with 869,556 annotations. These annotations were obtained from the semi-automatic labeling (see above) and include details such as the predicted vocalization type. The content of each column in the annotation file is described in Table 2.1. Each annotation corresponds to a single vocalization in one file. Most files include a single detection, though some files contain several vocalizations. 215,000 (72 hours) of these annotations were identified as a specific type of vocalization (see Figure 2.1 for the latent projection of all the vocalizations, colored by label).
3. One metadata file: Metadata.pdf, details the subjects and annotation definitions (Table 2.1, Supplementary Table S1).
4. An example of a raw audio file that is 5 minutes long.
5. A set of audio example files.

6. A sample Python code exemplifies data loading and plotting of vocalization spectrograms.
7. A sample Python code is exemplifying classifier loading and vocalization type prediction.

Column name	Description
filename	Name of the .wav file containing the vocalization, using the format Type_ID.wav
folder	Name of the folder containing the vocalization file, using the format YYYY_MM_folder ID
year	Start year of vocalization.
month	Start month of vocalization.
day	Start day of vocalization.
hour	Start hour of vocalization (since 2022)
minute	Start minute of vocalization (since 2022)
second	Start second of vocalization (since 2022)
millisecond	Start millisecond of vocalization (since 2022)
duration	Length of the vocalization file in seconds
recording file onset	Vocalization start time in the recording file (seconds)
recording file offset	Vocalization end time in the recording file (seconds)
type	Type of vocalization as classified by the model: Phee, Trill, Seep, Twitter, Tsik, Infant cry, or Vocalization by default
confidence	Confidence of the model in its type attribution (between 0 and 1)

Table 2.1: Annotation details. Descriptions of each column of the annotation file.

The recorded audio files are divided into folders by month of recording, with no more than 10,000 files per folder. The annotation and metadata files are in tabular separated-value format (TSV) to ease their use with automatic tools and allow direct upload into spreadsheet software. The metadata file includes descriptions of all identifiers in the annotation file. The example files contain several audio recordings that illustrate different recorded sounds. They are provided to help users become more familiar with the recorded data. These examples include Phee calls, Twitter calls, Infant cries, and examples of background noises.

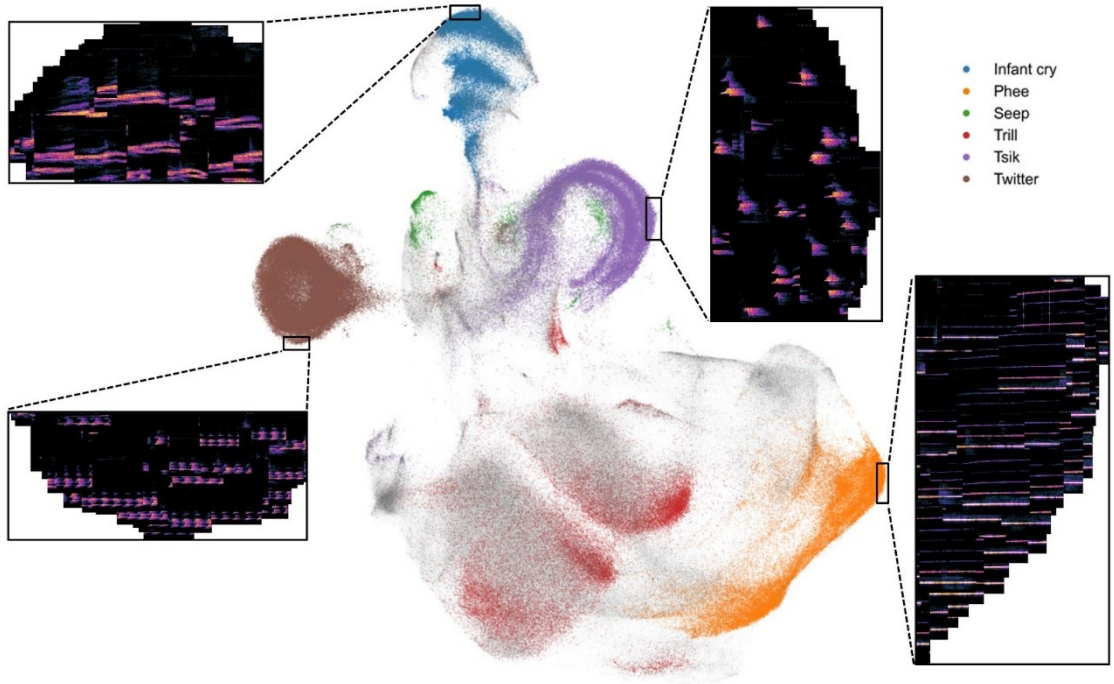


Figure 2.1: Latent projections of vocalizations. For each segmented vocalization, we computed a spectrotemporal representation. Using the trained encoder, we transformed these representations into a 16-dimensional space. We employed the UMAP technique from there to map the data into latent feature spaces. The colored points denote the predictions where the classifier assigned a high confidence score.

Since the dataset captures the specific times each vocalization was uttered, it paves the way for future research into the sequential organization of the marmoset vocal repertoire (see Figure 2.2 for a visual representation of the vocalization's temporal distribution; see Supplementary Table S4 for the distribution by label).

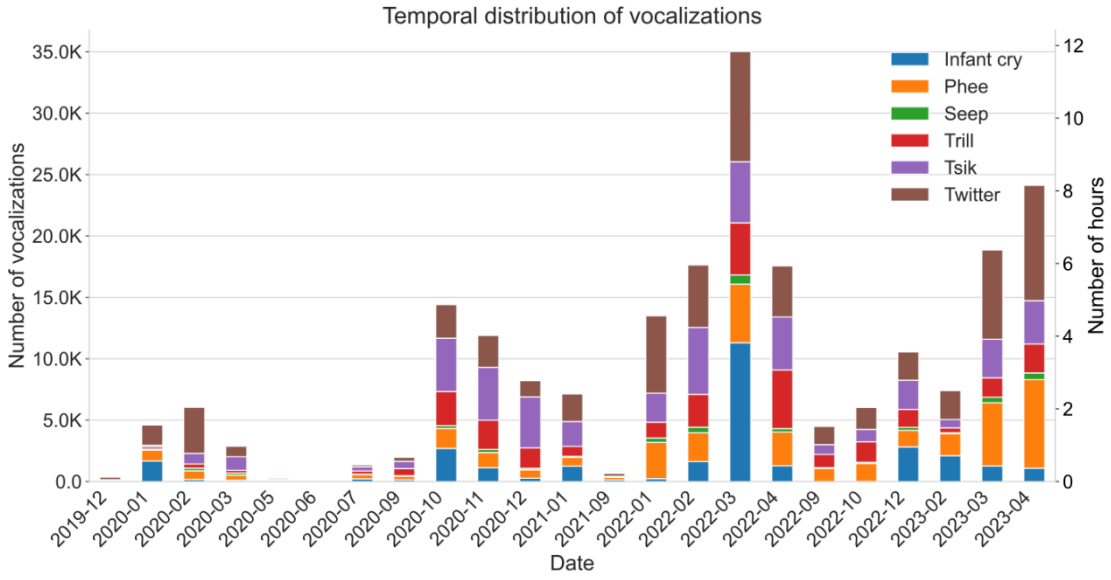


Figure 2.2: Temporal distribution of vocalizations over time. Distribution over time of 215,000 labeled vocalizations (72 hours in total). Each month, the proportion of vocalization type is indicated in thousands of vocalizations and hours. The proportion of labeled/labeled vocalization is 25/75% (unlabeled omitted here).

3.2. Code availability

The code is available on https://github.com/swasun/MarmAudioDataset.

3.3. Usage notes

First, you need to decompress the FLAC files:

```
python marmaudio/decompress_flac.py --folder_path=audios_compressed
```

Below is a short Python example that demonstrates how to load a wavefile based on some annotations found in 'Annotations.tsv':

```
from marmaudio.utils import read_waveform, denoise_waveform
import pandas as pd
import os

df = pd.read_csv('Annotations.tsv', sep='\t') # Read the annotations
print(df.prediction_type.value_counts()) # Display the labels distribution

random_row = df.sample(n=1) # Randomly sample a line for example purpose

file_path = os.path.join('marmaudio', 'audios',
f'{random_row["year"]}/{random_row["month"]}/{random_row["folder_id"]}',
f'{random_row["file_id"]}.wav')

signal, sampling_rate = read_waveform(file_path) # Read the vocalization
waveform and store it as 'signal'
```

```
denoised_signal = denoise_waveform(signal) # Denoise the signal if needed
```

Below is a short Python example that demonstrates how to load our pre-trained classifier and run it to predict the vocalization type of a loaded waveform:

```
from marmaudio.classifier import load_classifier, prediction_to_str

# signal = ... code to load a signal
clf = load_classifier()
prediction = clf(signal)
print(prediction_to_str(prediction))
```

4. Methods

4.1. Animal retrievals and cares

This study involved a total of thirty-five common marmosets (*Callithrix jacchus*) belonging to a colony of three families. Not all animals were present during the same period. The monkeys were not present for the entire data collection, notably due to conflicts or deaths. For more details on the periods of inclusion of each monkey, refer to Supplementary Table S1.

All animals included are the offspring of parents and grandparents born and raised in captivity for research purposes. All experimental procedures were in compliance with the European directive (2010/63/UE) and were approved by the Ethics Board of Institut de Neurosciences de la Timone (reference 2019010911313842).

4.2. Experimental setup

Acoustic recorders were set up in a lab with captive marmosets (Figure 2.3). The recordings were made using one microphone (C-100, Sony Corporation, Japan) placed directly in the room of three marmoset families (e.g., Supplementary Table S1) housed in cages (1.05 m long x 0.85 m wide x 2 m high). The mixing desk (RME Fireface UFX II, RME, Germany) and the computer allowing the recording via Adobe Audition (Adobe, CA, USA) were located in an adjacent room. Husbandry and technical rooms are soundproofed from the rest of the laboratory animal facility. Audio data was recorded from December 2019 to April 2023, consisting of 997 hours of data recording.

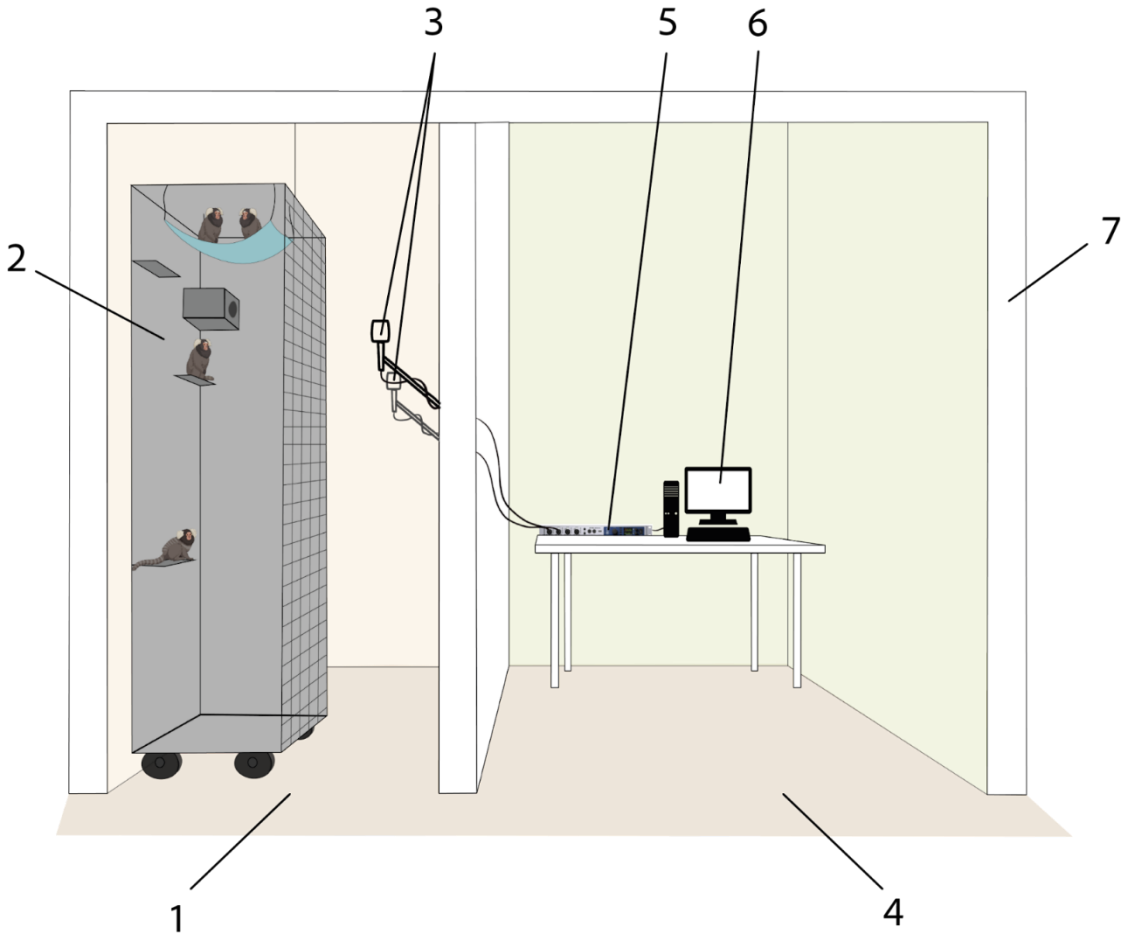


Figure 2.3: Schematic of the recording system. The diagram shown here is a schematic drawing of the recording setup, and the relative sizes and positions of the components are not to scale. The husbandry room (1) contained three cages (only one visible here, (2)) and two microphones (3). The technical room (4) was separated by a wall and contained a mixing desk (5) and a computer (6), allowing the recording. Husbandry and technical rooms were soundproof thanks to specialized insulation (7).

4.3. Segmentation and labeling

To build a dataset of marmoset vocalizations annotated by type, we followed the pipeline shown in Figure 2.4. Each step is described in this section.

To isolate vocalizations from background noise, we used a stationary noise reduction algorithm relying on spectral gating (*noisereduce* Python package; Sainburg et al., 2020). We then partially identified (recordings from 2019-2020) the vocalization sound events using a dynamic-thresholding segmentation algorithm (Sainburg et al., 2020), leading to 100,000 segmented audio events. The elements were partitioned in a spectro-temporal

manner, allowing for temporal overlap but ensuring frequency exclusivity between two elements (Figure 2.4, blue panel; see hyperparameters in Supplementary Table S2).

Given the large number of utterances to label, we opted for a semi-automated procedure leveraging unsupervised and self-supervised machine learning strategies to explore the sound event space and label the vocalization types, as well as filter out the noisy sound events (Figure 2.4, orange panel). A convolutional autoencoder (network architecture and particularities of the training procedure are detailed in Best et al., 2023) was trained on segmented time-frequency representations of 0.5 seconds to encode them into a 16-dimensional latent space allowing the measurement of vocalization similarity (Sainburg et al., 2020; Best et al., 2023). The representations were Mel-spectrograms (short-time Fourier Transform (STFT) with a Hann window of 1,024, no FFT padding, and a hop size of 368), with Mel filterbank of 128 bands between 1 kHz and 48 kHz. The Mel scale is a popular choice of center frequencies aiming to mimic pitch perception characteristics of the human auditory system. These representations were subsequently treated as points in a feature space after applying the dimensionality reduction algorithm UMAP (McInnes et al., 2018). We then clustered vocalizations close to one another in feature space, using a density-based algorithm (McInnes et al., 2017), allowing the annotation of vocalizations by type (Figure 2.4, orange panel, ‘Clustered sound events’). Clusters, which encompass hundreds to thousands of sound events, were meticulously examined by experts. They associated these clusters with specific call types and filtered out any misclassifications. For each cluster, an expert reviewed a folder of spectrogram images, discarding any that did not align with the cluster's general trend. Subsequently, these cluster sounds were categorized by ‘vocalization’ type or as 'noise.' This process yielded a partially labeled database, essential for the subsequent iterative label refinement procedure.

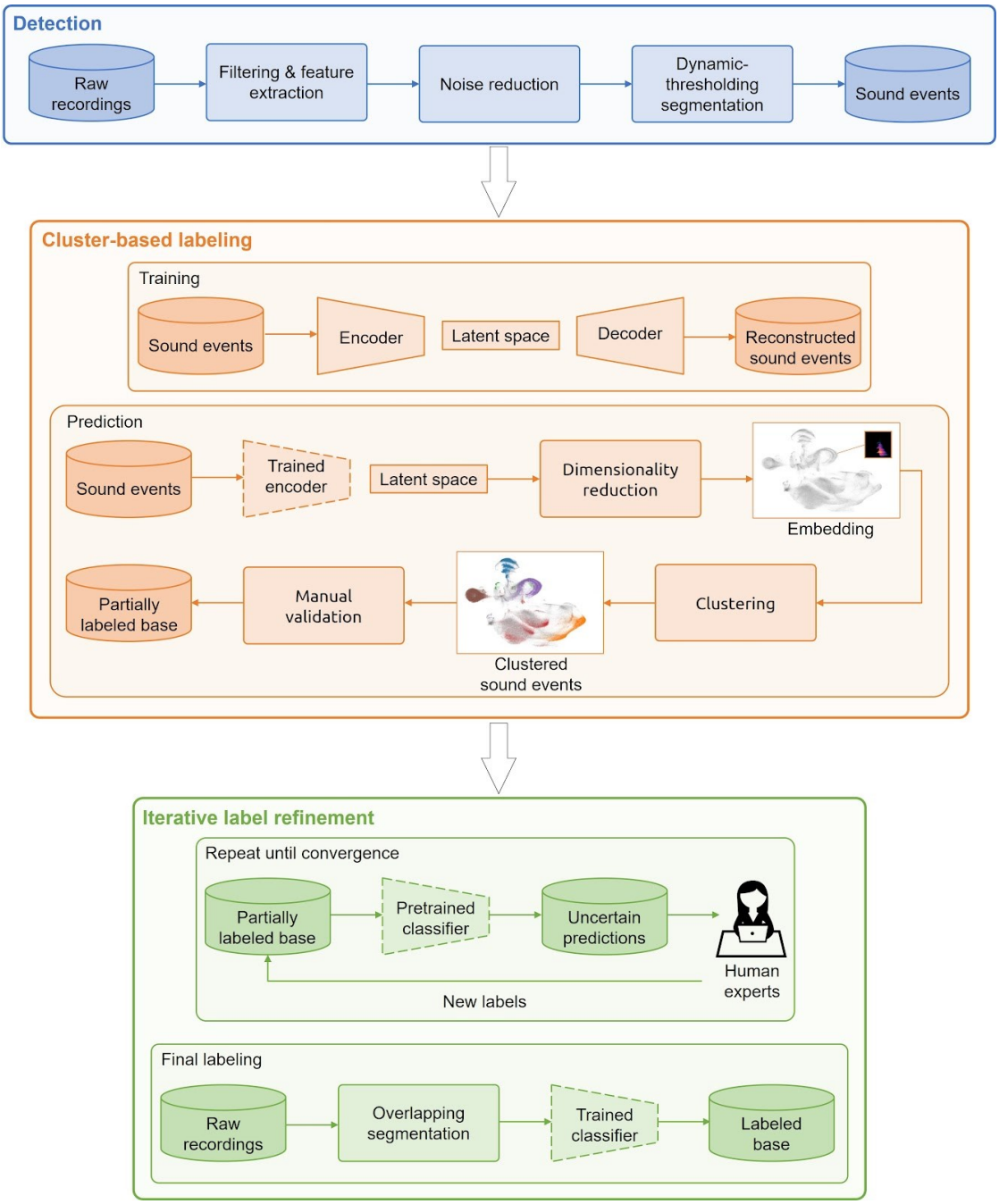


Figure 2.4: Pipeline for the creation of the published database.

After compiling the initial database, we engaged in an iterative process: We trained a classifier and then improved its predictions by visually inspecting and manually correcting multiple spectrograms displayed simultaneously. These spectrograms were sampled based on mislabels with high confidence (Figure 2.4., green panel). We continued this process until the classifier's performance met a threshold of 0.7, which was found empirically. We introduced custom thresholds for each label type to refine the classifier's decisions based on the prediction confidence, thus optimizing the label assignments.

These thresholds were identified based on the empirical observations from our dataset and reflect the distinctive nature of each label type (Phee ≥ 0.7; Seep ≥ 0.86; Trill ≥ 0.86; Tsik ≥ 0.7; Twitter ≥ 0.7). Any vocalization with prediction confidence below these label-specific thresholds was accordingly re-labeled as 'Vocalization' (i.e., a vocalization of unknown type). We adjusted each sound event's start and end times based on its predicted label post-classification (Supplementary Table S3).

4.4. Technical validation

The annotation types were defined by Dr. Manon Obliger-Debouche and Dr. Sabrina Ravel. The recordings were annotated semi-automatically by myself and Dr Paul Best. These observers were certified after annotating a few recording days, which were then validated by an expert (Manon Obliger-Debouche or Sabrina Ravel). In annotating the recordings, we adopted a conservative approach, in which we designated as 'unknown' any data for which we had any doubt. Despite the training of the observers, some noise might have been introduced during the manual annotations and by the annotating algorithms. Thus, we estimated an error rate by a post-hoc quality test (procedure from Prat et al., 2017): 700 annotated recordings (100 per label type) were sampled randomly and were then carefully re-annotated by Manon Obliger-Debouche, myself, Paul Best, and Sabrina Ravel. Errors were counted when there was a discrepancy between the post-hoc and the original annotations or when the post-hoc examination concluded that some doubt still existed (e.g., if only 3 out of the 4 confirm it, it is considered an error). The error rates were, on average, 9.43% (90.57% Confidence-Interval [CI]: 86.00–95.00%) for the vocalization type identification (see Table 2.2 below for scores per label type).

Label	Error rate (%)	Accuracy (%)	CI lower (%)	CI upper (%)
Infant cry	1.00	99.00	97.00	100.00
Phee	0.00	100.00	100.00	100.00
Seep	21.00	79.00	71.00	87.00
Trill	18.00	82.00	74.00	90.00

Tsik	16.00	84.00	77.00	91.00
Twitter	2.00	98.00	95.00	100.00
Vocalization	8.00	92.00	87.00	97.00
Average	9.43	90.57	86.00	95.00

Table 2.2: Vocalization type error rates. 700 recordings (100 per label type) were re-reviewed by 4 experts. Errors were noted for inconsistencies or lingering uncertainties.

5. Conclusion

This chapter addressed the utilization of computational methods in studying the evolution of vocal communication among primates. We highlighted the need for more comprehensive datasets for primate vocalizations, specifically for macaque and marmoset monkeys. We underscored its significance in understanding the coding principles in voice patches across primate species.

We presented an end-to-end pipeline for extracting and analyzing vocalizations from marmoset monkey recordings, continuously recorded with a sampling rate of 96 kHz from a stabulation room housing ~20 marmosets in three cages simultaneously. The dataset comprises over 800,000 files, amounting to 253 hours of data collected over 40 months. Each recording lasts a few seconds and captures the marmosets' social vocalizations, encompassing their entire known vocal repertoire during the experimental period. Around 215,000 calls were annotated with the vocalization type. The provided dataset, source code, and pre-trained classifier offer a resource for future research in this domain. Moving forward, it is crucial to expand upon this foundational work by incorporating more species-specific vocal datasets and refining computational methodologies to further our understanding of the evolution of vocal communication. This point and others are discussed in Section 3 of the Discussion chapter.

Chapter 3

Encoding and decoding of voice identity in human auditory cortex

Chapter 1 discussed the cerebral processing of voice information in humans and non-human primates. It is established that conspecific vocalizations preferentially activate the "temporal voice areas" (TVAs). However, how these areas represent voice identity information—such as speaker gender and specific identity—remains unclear. This chapter examines the correlation between brain activity and voice identity, as measured using neuroimaging techniques and representations derived from deep learning. We conduct computational experiments—including neural encoding, neural decoding, and representational similarity analysis (RSA)—to bridge the deep learning-based voice representation with the fMRI responses to voice stimuli.

1. Abstract

The cerebral processing of voice information is known to engage, in human as well as non-human primates, "temporal voice areas" (TVAs) that respond preferentially to conspecific vocalizations. However, how voice information is represented by neuronal populations in these areas, particularly speaker identity information, remains poorly understood. Here, we used a deep neural network (DNN) to generate a high-level, small-dimension representational space for voice identity—the 'voice latent space' (VLS)—and examined its linear relation with cerebral activity via encoding, representational similarity, and decoding analyses. We find that the VLS maps onto fMRI measures of cerebral activity in response to tens of thousands of voice stimuli from hundreds of different speaker identities and better accounts for the representational geometry for speaker identity in the TVAs than in A1. Moreover, the VLS allowed TVA-based reconstructions of voice stimuli that preserved essential aspects of speaker identity as assessed by both machine classifiers and human listeners. These results indicate that the DNN-derived VLS provides high-level representations of voice identity information in the TVAs.

2. Introduction

In recent years, deep neural networks (DNNs) have emerged as a powerful tool for representing complex visual data, such as images (LeCun et al., 2015) or videos (Liu et al., 2020). In the auditory domain, DNNs have been shown to provide valuable representations—so-called feature or latent spaces—for modeling the cerebral processing of sound (brain encoding) (speech: Kell et al., 2018; Millet et al., 2022; semantic content: Caucheteux et al., 2022; Caucheteux et King, 2022; Caucheteux et al., 2023; Giordano et al., 2023; music: Güçlü et al., 2016), or reconstructing the stimuli listened by a participant (brain decoding) (Akbari et al., 2019). They have not yet been used to explain cerebral representations of identity-related information due in part to the focus on speech information (von Kriegstein, 2003; Morillon et al., 2022).

Here, we addressed this challenge by training a 'Variational autoencoder' (VAE; Kingma et Welling, 2014) DNN to reconstruct voice spectrograms from 182,000 250-ms voice samples from 405 different speaker identities in 8 different languages from the CommonVoice database (Ardila et al., 2020). Brief (250 ms) samples were used to

emphasize speaker identity-related information in voice, already available after a few hundred milliseconds (Schweingerger et al., 1997; Lavan, 2023), over linguistic information unfolding over longer periods. While a quarter of a second is admittedly short compared to standards of, e.g., computational speaker identification that typically uses 2-3s samples, this short duration is sufficient to allow near-perfect gender classification and performance levels well above chance for speaker discrimination (Figure 3.5). This brief duration allowed the presentation of many more stimuli to our participants in the scanner while preserving acceptable behavioral and classifier performance levels.

State-of-the-art studies have primarily relied on task-optimized neural networks (i.e., DNN trained using supervised learning to classify a category from the input) to study sensory cortex processes (Yamins et DiCarlo, 2016; Schrimpf et al., 2018). They can reach high accuracies in brain encoding (Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Han et al., 2019). However, there is increasing evidence that unsupervised learning, such as that used for the VAE, also provides plausible computational models for investigating brain processing (Higgins et al., 2021; Zhuang et al., 2021; Millet et al., 2022; Orhan et al., 2022). Thus, the VAE-derived VLS, exploited within encoding, representational similarity, and decoding frameworks, offers a potentially promising tool for investigating the representations of voice stimuli in the secondary auditory cortex (Naselaris et al., 2011). Autoencoders learn to compress stimuli with high dimensionality into a lower-dimensional space that nonetheless allows reconstruction of the original stimuli via an inverse transformation learned by the second part of the network called the decoder. Figure 3.1a shows the architecture of the VAE, with its encoder that reduces an input spectrogram to a highly compressed, 128-dimension *voice latent space* (VLS) representation and its decoder that reconstructs the spectrogram from this VLS representation. Points in the VLS correspond to voice samples with different identities and phonetic content. A line segment in the VLS contains points corresponding to perceptual interpolations between its two extremities (Figure 3.1b; Supplementary Audio S1). VLS coordinates of samples presented to the participants averaged by speaker identity suggest that a major organizational dimension of the latent space is voice gender (Figure 3.1b) (colored by age or language in Supplementary Figure S1).

In order to test whether VLS accounts well for cerebral activity in response to voice stimuli, we scanned three healthy volunteers using fMRI to measure an indirect index of their cerebral activity across 10+ hours of scanning each in response to ~12,000 of the voice samples, denoted *BrainVoice* in the following, used to train the DNN. The small number of participants does not allow for generalization at the general population level as in standard fMRI studies. However, it allows testing for replicability as in comparable studies involving 10+ hours of scanning per participant (VanRullen & Reddy, 2019). Different stimulus sets were used across participants to provide a stringent test of replicability based on subject-level analyses. Stimuli consisted of randomly spliced 250-ms excerpts of speech samples from the CommonVoice database (Ardila et al., 2020) by 119 speakers in 8 languages. For assessing generalization performances of decoding models and brain-based reconstruction, six test stimuli were repeated more often (60 times) for each participant to provide robust estimates of their induced cerebral activity (see Methods). We first modeled these responses to voice using a general linear model (GLM) (Friston et al., 1994) with several nuisance regressors as an initial denoising step (Supplementary Figure S4), then used a second GLM modeling cerebral responses to the different speaker identities (Supplementary Figure S3a), resulting in one voxel activity map per speaker (Supplementary Figure S3b). We independently localized in each participant several regions of interest (ROIs) on which subsequent analyses were focused: the anterior, middle and posterior TVAs in each hemisphere (individually localized via an independent 'voice localizer scan' and MNI coordinates provided in Pernet et al., 2015; Supplementary Figure S3c) as well as primary auditory cortex (A1) (using a probabilistic map in MNI space (Penhune et al., 1996; Supplementary Figure S3d).

We first asked how the VLS could account for the brain responses to speaker identities (encoding) measured in A1 and the TVAs compared to a linear autoencoder's latent space (LIN). This approach was chosen because it has been demonstrated that a linear autoencoder with a d-dimensional hidden layer projects data in the same subspace as the one spanned by the d first eigenvectors of a principal component analysis (PCA) (Plaut et al., 2018). We used a general linear model (GLM) of fMRI responses to the speaker identities, resulting in one voxel activity map per speaker (Supplementary Figure S3). Then, we computed the average VLS coordinates of the fMRI voice stimuli for each speaker identity, which may be seen as a speaker representation in the VLS (see *Identity-based and*

stimulus-based representations section). Next, we trained a linear voxel-based encoding model to predict the speaker voxel activity maps from the speaker VLS coordinates. As VAE achieves compression through a series of nonlinear transformations (Wetzel, 2017), we contrast its results with a linear autoencoder's latent space. This method has previously been applied to fMRI-based image reconstructions (Cowen et al., 2014; VanRullen & Reddy, 2019; Mozafari et al., 2020).

The extent to which the VLS allows linearly predicting the fMRI recordings does not provide insight into the representational geometries, i.e., the differences between the patterns of cerebral activity for speaker identity. We addressed this question by using representational similarity analysis (RSA; Kriegeskorte et al., 2008) to test which model better accounts for the representational geometry for voice identities in the auditory cortex. Using RSA as a model comparison framework is relevant to examining the brain-model relationship from complementary angles (Diedrichsen et al., 2017). We built speaker x speaker representational dissimilarity matrices (RDMs) capturing pairwise differences in cerebral activity or model predictions between all pairs of speakers; then, we examined how well the LIN and VLS-derived RDMs correlated with the cerebral RDMs from A1 and the TVAs.

A robust test of the adequacy of models of brain activity, and a long-standing goal in computational neurosciences, is the reconstruction of a stimulus presented to a participant from the evoked brain responses. While reconstruction of visual stimuli (images, videos) from cerebral activity has been performed by a number of groups (VanRullen et Reddy, 2019; Mozafari et al., 2020; Le et al., 2022; Gaziv et al., 2022; Dado et al., 2022; Chen et al., 2023), validating the DNN-derived representational spaces, comparable work in the auditory domain is scarce, almost exclusively concentrated on linguistic information (Santoro et al., 2017). Akbari et al. used a DNN to reconstruct speech stimuli based on ECoG recording of auditory cortex activity, an invasive method compared to techniques like fMRI. They obtained a good phonetic recognition rate but chance-level gender categorization performance from reconstructed spectrograms and no evaluation of speaker identity discrimination.

Here, we built on the linear relationship uncovered in our encoding analysis between the VLS and the fMRI recordings to invert it and try to predict VLS coordinates from the recorded fMRI data; then, using the decoder, we reconstructed the spectrograms of stimuli presented to the participants (Wu et al., 2006; Naselaris et al., 2011). The voice identity information available in the reconstructed stimuli was finally assessed by human listeners using both machine learning classifiers and behavioral tasks (Figure 3.4).

3. Results

3.1. Voice Information in the Voice Latent Space (VLS)

In order to probe the informational content of the VLS, linear classifiers were trained to categorize the voice stimuli from 405 speakers by gender (2 classes), age (2 classes) or identity (119 classes, cf Methods) based on VLS coordinates, or their LIN features as control (Figure 3.1c,d,e; we aggregated the stimuli from the 3 participants; for each model computed the latent space of each stimulus and averaged the latent spaces by speaker identity, leading to 405 128-dimensional vectors. We then trained linear classifiers using a 5-fold cross-validation scheme, see *Characterization of the autoencoder latent space*). The mean of the distribution of accuracies obtained for 100 random classifier initializations (as to account for variance; Bouthillier et al., 2021) was significantly above chance level (all ps < 1e-10) for all classifications (LIN: gender (mean accuracy ± s.d.) = 97.64±1.77%, t(99)=266.94; age: 64.39±4.54%, t(99)=31.53; identity: 40.52±9.14%, t(99)=39.37; VLS: gender: 98.59±1.19%, t(99)=406.47; age: 67.31±4.86%, t(99)=35.41; identity: 38.40±8.75%, t(99)=38.73). We then evaluated the difference in performance at preserving identity-related information between the VLS and LIN via one-way ANOVAs. Results showed a significant effect of Feature (LIN/VLS) in categories (all Fs(1, 198) > 225.15, all ps<.0001) but not in identity. Post-hoc paired t-tests showed that the VLS was better than the LIN at encoding information related to voice identity, as evidenced by a significant difference in means for gender (t(99)=-6.11, p<.0001), age (t(99)=-6.10, p<.0001) but not for identity classifications (t(99)=1.71).

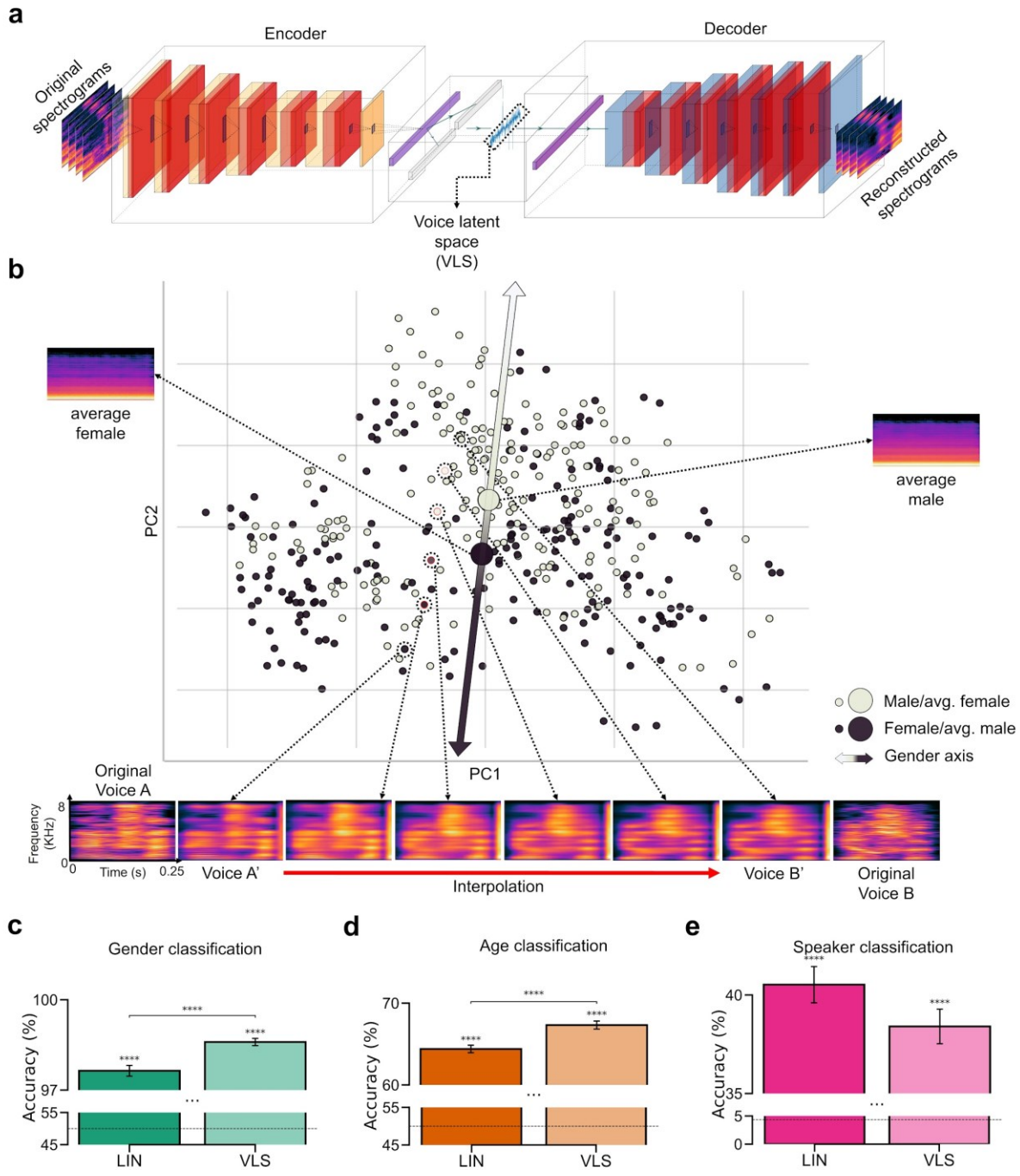


Figure 3.1: DNN-derived Voice Latent Space (VLS). a, Variational autoencoder (VAE) Architecture. Two networks learned complementary tasks. An encoder was trained using 182K voice samples to compress their spectrogram into a 128-dimension representation, the voice latent space (VLS), while a decoder learned the reverse mapping. The network was trained end-to-end by minimizing the difference between the original and reconstructed spectrograms. **b,** Distribution of the 405 speaker identities along the first 2 principal components of the VLS coordinates from all sounds, averaged by speaker identity. Each disk represents a speaker's identity colored by gender. PC2 largely maps onto voice gender (ANOVAs on the first two components: PC1: $F(1, 405)=0.10$, $p=.74$; PC2: $F(1, 405)=11.00$, $p<.001$). Large disks represent the

average of all male (black) or female (gray) speaker coordinates, with their associated reconstructed spectrograms (note the flat fundamental frequency (f_0) and formant frequencies contours caused by averaging). The bottom of the spectrograms illustrates an interpolation between stimuli of two different speaker identities: spectrograms at the extremes correspond to two original stimuli (A, B) and their VLS-reconstructed spectrograms (A', B'). Intermediary spectrograms were reconstructed from linearly interpolated coordinates between those two points in the VLS (red line) (cf. Supplementary Audio S1). **c,d e,** Performance of linear classifiers at categorizing speaker gender (chance level: 50%), age (young/adult, chance level: 50%), or identity (119 identities, chance level: 0.84%) based on VLS or LIN coordinates. Error bars indicate the standard error of the mean (s.e.m) across 100 random classifier initializations. All ps<1e-10. The horizontal black dashed lines indicate chance levels. ****: p<0.0001.

Thus, despite its low number of dimensions (each input spectrogram has 401x21=8421 parameters and is summarized in the VLS by a mere 128 dimensions), the VLS appears to meaningfully represent the different sources of voice information perceptually available in the vocal stimuli. This representational space, therefore, constitutes a relevant candidate for linearly modeling voice stimulus representations by the brain.

3.2. Brain Encoding

We used a linear voxel-based encoding model to test whether VLS linearly maps onto cerebral responses to speaker identities measured with fMRI in the different ROIs. A regularized linear regression model (cf. Methods) was trained on a subset of the data (5-fold cross-validation scheme) to predict the voxel maps for each speaker identity. For each fold, the trained model was tested on the held-out speaker identities (Figure 3.2a). The model's performance was assessed for each ROI using the Pearson correlation score between each voxel's actual and predicted responses (Schrimpf et al., 2021). Similar predictions were tested with features derived from LIN (cf. Methods). Figure 3.2b shows the distribution of correlation coefficients obtained for each of the ROIs for the 2 sets of features across voxels, hemispheres, and participants.

One-sample t-tests showed that the means of Fisher z-transformed coefficients for both LIN features and VLS were significantly higher than zero (LIN: A1 t(197)=7.25, p<.0001, pTVA t(175)=4.49, p<.0001, mTVA t(164)=9.12, p<.0001 and aTVA t(147)=6.81, p<.0001;

VLS: A1 t(197)=4.76, p<.0001, mTVA t(164)=10.12, p<.0001 and aTVA t(147)=5.52, p<.0001 but not pTVA t(175)=-1.60) (Supplementary Tables 2-3).

A mixed ANOVA performed on the Fisher z-transformed coefficients with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors showed a significant effect of Feature (F(3, 683)=56.65, p<.0001), a significant effect of ROI (F(3, 683)=18.50, p<.0001), and a moderate interaction Feature x ROI (F(3, 683)=5.25, p<.01). Post-hoc comparisons revealed that the mean of correlation coefficients was higher for LIN than for VLS in A1 (t(197)=4.02, p<.0001), pTVA (t(175)=6.64, p<.0001), aTVA (t(147)=3.78, p<.001) but not in mTVA (t(164)=0.58) (Supplementary Table 4); and that the voxel patterns are better predicted in mTVA than in A1 for both models (LIN: t(361)=2.36, p<.05); VLS: t(361)=4.91, p<.0001) (Supplementary Table 5). However, inspecting the distribution of model-voxel correlations, we found that both models account for different parts of the voice identity responses and differ across ROIs (Figure 3.2c).

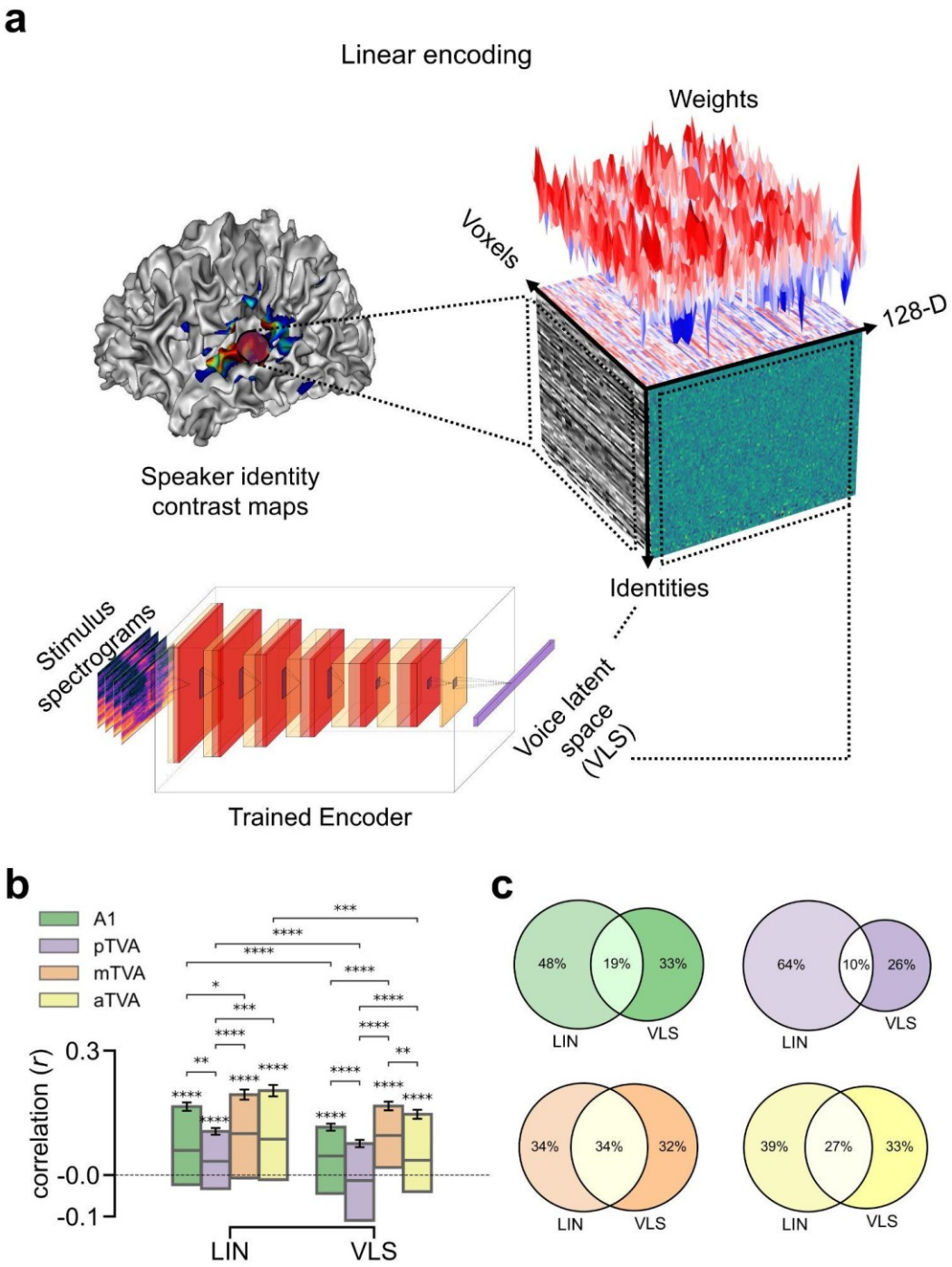


Figure 3.2: Predicting brain activity from the VLS. a, Linear brain activity prediction from VLS for ~135 speaker identities in the different ROIs. We first fit a GLM to predict the BOLD responses to each voice speaker identity. Then, using the trained encoder, we computed the average VLS coordinates of the voice stimuli presented to the participants based on speaker identity. Finally, we trained a linear voxel-based encoding model to predict the speaker voxel activity maps from the speaker VLS coordinates. The cube illustrates the linear relationship between the fMRI responses to speaker identity and the VLS coordinates. The left face of the cube represents the activity of the voxels for each speaker's identity, with each line corresponding to one speaker. The

right face displays the VLS coordinates for each speaker's identity. The cube's top face shows the encoding model's weight vectors. **b,** Encoding results. For each region of interest, the model's performance was assessed using the Pearson correlation score between the true and the predicted responses of each voxel on the held-out speaker identities. Pearson's correlation coefficients were computed for each voxel on the speakers' axis and then averaged across hemispheres and participants. Similar predictions were tested with the LIN features. Error bars indicate the standard error of the mean (s.e.m) across voxels. *p < 0.05; **p < 0.01; **p < 0.001; ****p < 0.0001. **c,** Venn diagrams of the number of voxels in each ROI with the LIN, the VLS, or both models. For each ROI and each voxel, we checked whether the test correlation was higher than the median of all participant correlations (intersection circle), and if not, which model (LIN or VLS) yielded the highest correlation (left or right circles).

3.3. Representational Similarity Analysis

For RSA, we built speaker x speaker representational dissimilarity matrices (RDMs), capturing for each ROI the dissimilarity in voxel space between each pair of speaker voxel maps ('brain RDMs'; cf. Methods) using Pearson's correlation (Walther et al., 2016). We compared these four bilateral brain RDMs (A1, aTVA, mTVA, pTVA) to two 'model RDMs' capturing speaker pairwise feature differences predicted by LIN and the VLS (Figure 3.3a) built using cosine distance (Xing et al., 2015; Bhattacharya et al., 2017; Wang et al., 2018). Figure 3.3b shows for each ROI the Spearman correlation coefficients between the brain RDMs and the two model RDMs for each participant and hemisphere (Kriegeskorte et al., 2008; Figure 3.3c for an example of brain-model correlation).

These brain-model correlation coefficients were compared to zero using a 'maximum statistics' approach based on random permutations of the model RDMs' rows and columns (Maris & Oostenveld, 2007; cf. Methods; Figure 3.3b). For the LIN model, only one brain-model RDM correlation was significantly different from zero (one-tailed test): in mTVA, right hemisphere in S3 (p=.0500). For the VLS model, in contrast, 5 significant brain-model RDM correlations were observed in all four ROIs: in A1, right hemisphere in S3 (p=.0142); pTVA: right hemisphere in S3 (p=.0160); mTVA: left hemisphere in S3 (p=.007); aTVA: left hemispheres in S1 (p=.0417) and S3 (p=.0001) (Supplementary Table 6).

A two-way repeated-measures ANOVA with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors performed on the Fisher z-transformed correlation coefficients showed a tendency towards a significant effect of Feature ($F(1, 2)=22.53$, $p=.04$), and no ROI ($F(3, 6)=1.79$, $p=.30$) or interaction effects ($F(3, 6)=1.94$, $p=.22$). We compared the correlation coefficients between the VLS and LIN models within participants and hemispheres using one-tailed tests, based on the a priori hypothesis that the VLS models would exhibit greater brain-model correlations than the LIN models (cf. Methods). The results revealed two significant differences in one of the three participants, both favoring the VLS model (S3: right pTVA, $p=.0366$; left aTVA, $p=.00175$) (Supplementary Table 7).

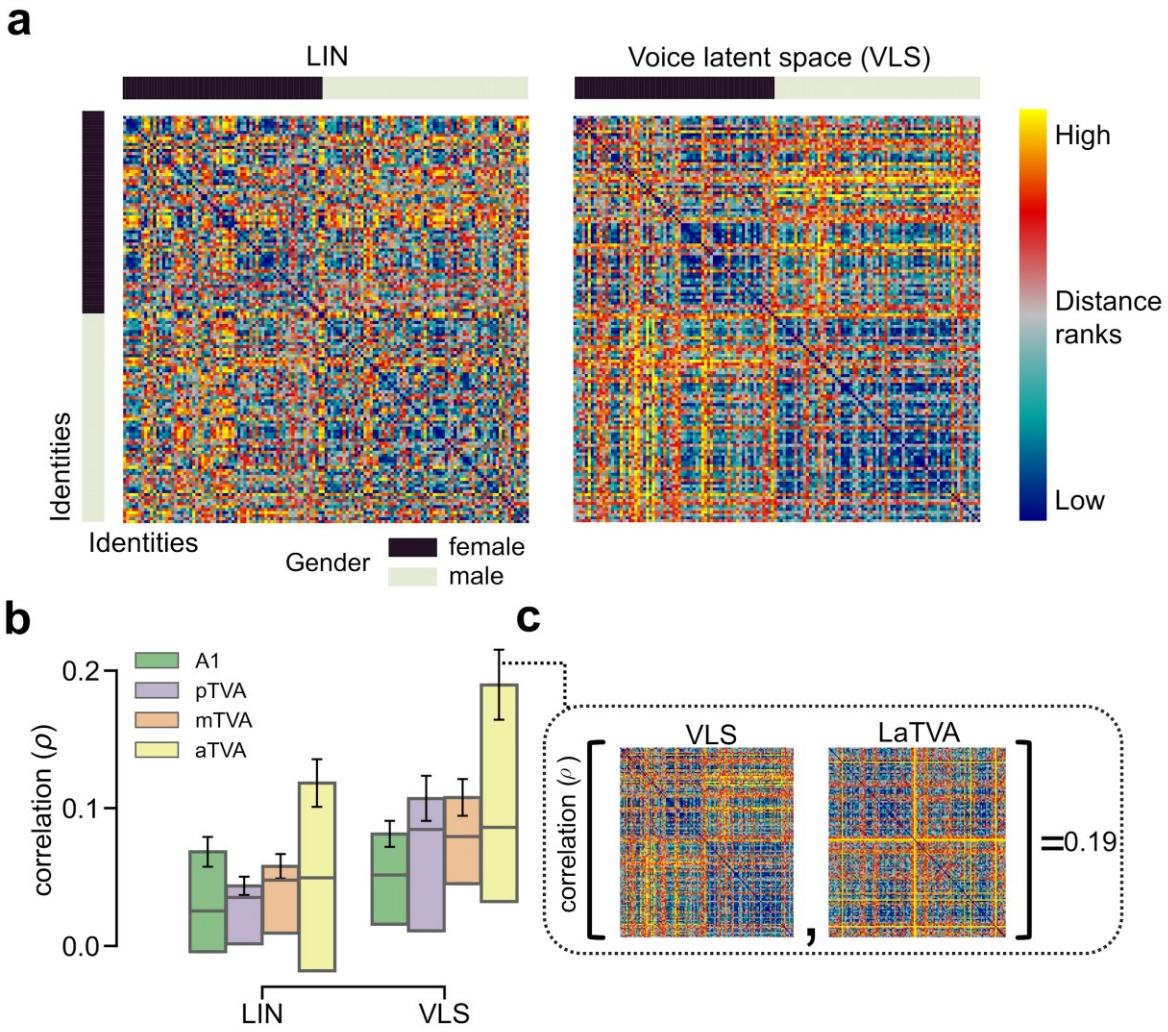


Figure 3.3: The VLS better explains representational geometry for voice identities in the TVAs than the linear model. a, Representational dissimilarity matrices (RDMs) of pairwise speaker dissimilarities for ~135 identities (arranged by gender, cf. sidebars), according to LIN and VLS. **b,** Spearman correlation coefficients between the brain RDMs for A1, the 3 TVAs, and the

2 model RDMs. Error bars indicate the standard error of the mean (s.e.m) across brain-model correlations. **c,** Example of brain-model RDM correlation in the TVAs. The VLS RDM and the brain RDM yielding one of the highest correlations (LaTVA) are shown in the insert.

3.4. Decoding and Reconstruction

We finally inverted the brain-VLS relationship to predict linearly VLS coordinates based on fMRI measurements (Figure 3.4a; see 'Brain decoding' in Methods) and reconstructed via the trained decoder the spectrograms of 18 Test Stimuli (3 participants x 6 stimuli per participant; see Figure 3.4b, and Supplementary Audio S2; audio estimated from spectrogram through phase reconstruction).

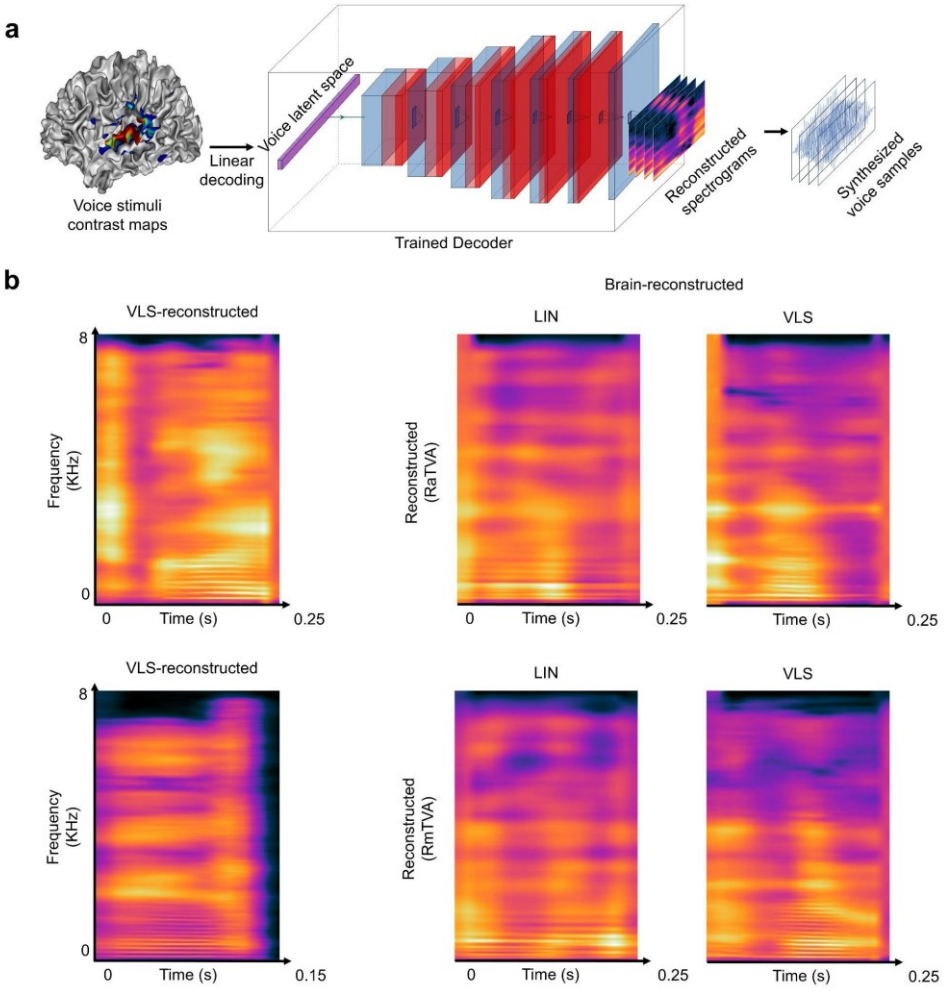


Figure 3.4: Reconstructing voice identity from brain recordings. a, A linear voxel-based decoding model was used to predict the VLS coordinates of 18 Test Stimuli based on fMRI responses to ~12,000 Train stimuli in the different ROIs. To reconstruct the audio stimuli from the brain recordings, the predicted VLS coordinates were then fed to the trained decoder to yield reconstructed spectrograms, synthesized into sound waveforms using the Griffin-Lim phase reconstruction algorithm (Griffin & Lim, 1983). **b,** Reconstructed spectrograms of the stimuli

presented to the participants. The left panels show the spectrogram of example original stimuli reconstructed from the VLS, and the right panels show brain-reconstructed spectrograms via LIN and the VLS (cf Supplementary Audio S2).

We first assessed the nature of the reconstructed stimuli by using a DNN trained to categorize natural audio events (Howard et al., 2017): all reconstructed versions of the 18 Test Stimuli were categorized as 'speech' (1 class out of 521 - no 'voice' classes). To evaluate the preservation of voice identity information in the reconstructed voices, pre-trained linear classifiers were used to classify the speaker gender (2 classes), age (2 classes), and identity (17 classes) of the 18 reconstructed Test Stimuli. The mean of the accuracy distribution obtained across random classifier initializations (20 per ROI) used on the stimuli reconstructed from the induced brain activity was significantly above chance level for gender (LIN: pTVA (mean accuracy ± s.d.): 72.08±5.48, t(39)=25.15; VLS: A1: 61.11±2.15, t(39)=32.25; pTVA: 63.89±2.78, t(39)=31.22), age (LIN: pTVA: 54.58±4.14, t(39)=6.90; aTVA: 63.96±12.55, t(39)=6.94; VLS: pTVA: 65.00±7.26, t(39)=12.89; aTVA: 60.42±5.19, t(39)=12.54) and identity (LIN: A1: 9.20±9.23, t(39)=2.24; pTVA: 9.48±4.90, t(39)=4.59; aTVA: 9.41±6.28, t(39)=3.51; VLS: pTVA: 16.18±7.05, t(39)=9.11; aTVA: 8.23±4.70, t(39)=3.12) (Figure 3.5a-c; Supplementary Tables 8-10).

Two-way ANOVAs with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors performed on classification accuracy scores (gender, age, identity) revealed for gender classifications significant effects of Feature $F(1, 312)=12.82$, $p<.0005$) and ROI (gender: $F(3, 312)=245.06$, $p<.0001$; age: $F(3, 312)=64.49$, $p<.0001$; identity: $F(3, 312)=14.49$, $p<.0001$), as well as Feature x ROI interactions (gender: $F(3, 312)=56.74$, $p<.0001$; age: $F(3, 312)=4.31$, $p<.001$; identity: $F(3, 312)=8.82$, $p<.0001$). Post-hoc paired t-tests indicated that the VLS was better than LIN in preserving gender, age and identity information in at least one TVA compared with A1 (gender: aTVA: $t(39)=5.13$, $p<.0001$; age: pTVA: $t(39)=9.78$, $p<.0001$; identity: pTVA: $t(39)=4.01$, $p<.0005$) (all tests in Supplementary Table 11). Post-hoc two sample t-tests comparing ROIs revealed significant differences in all classifications, in particular with pTVA outperforming other ROIs in gender (LIN: pTVA vs A1: $t(78)=22.40$, $p<.0001$; pTVA vs mTVA: $t(78)=10.92$, $p<.0001$; pTVA vs aTVA: $t(78)=31.47$, $p<.0001$; VLS: pTVA vs A1: $t(78)=4.94$, $p<.0001$;

pTVA vs mTVA: t(78)=13.96, p<.0001; pTVA vs aTVA: t(78)=22.06, p<.0001), age (LIN: pTVA vs A1: t(78)=7.26, p<.0001; pTVA vs mTVA: t(78)=10.11, p<.0001; VLS: pTVA vs A1: t(78)=5.71, p<.0001; pTVA vs mTVA: t(78)=10.11, p<.0001; pTVA vs aTVA: t(78)=3.21, p<.005) and identity (LIN: pTVA vs mTVA: t(78)=2.27, p<.05; VLS: pTVA vs A1: t(78)=6.45, p<.0001; pTVA vs mTVA: t(78)=6.62, p<.0001; pTVA vs aTVA: t(78)=5.85, p<.0001) (Supplementary Table 12).

We further evaluated voice identity information in the reconstructed stimuli by testing human participants (n=13) in a series of 4 online experiments assessing the reconstructed stimuli on (i) naturalness judgment, (ii) gender categorization, (iii) age categorization, and (iv) speaker categorization (cf Methods). The naturalness rating task showed that the VLS-reconstructed stimuli sounded more natural compared to LIN-reconstructed ones, as revealed by a two-way repeated-measures ANOVA (factors: Feature and ROI) with a strong effect of Feature (F(1, 12)=53.72, p<.0001) and a small ROI x Feature interaction (F(3, 36)=5.36, p<.005). Post-hoc paired t-tests confirmed the greater naturalness of VLS-reconstructed stimuli in both A1 and the TVAs (all ps<.0001) (Figure 3.5g). For the gender task, one-sample t-tests showed that categorization of the reconstructed stimuli was only significantly above chance level for the VLS (A1: (mean accuracy ± s.d.) 55.77±10.84, t(25)=2.66, p<.01; pTVA: 61.75±7.11, t(25)=8.26, p<.0001; aTVA: 55.13±9.23, t(25)=2.78, p<.01). Regarding the age and speaker categorizations, results also indicated that both the LIN- and VLS-reconstructed stimuli yielded above-chance performance in the TVAs (age: LIN: aTVA, 55.77±14.95, t(25)=1.93, p<.05; VLS: aTVA, 63.14±11.82, t(25)=5.56, p<.0001; identity: LIN: pTVA: 54.38±9.34, t(17)=1.93, p<.05; VLS: pTVA: 63.33±6.75, t(17)=8.14, p<.0001) (Supplementary Tables 13-15). Two-way repeated-measures ANOVAs revealed a significant effect of ROI for all categories (gender: F(3, 27)=5.90, p<.05; age: F(3, 36)=14.25, p<.0001; identity: F(3, 24)=38.85, p<.0001), and a Feature effect for gender (F(1, 9)=43.61, p<.0001) and identity (F(1, 8)=14.07, p<.001), but not for age (F(1, 12)=4.01, p=0.07), as well as a ROI x Feature interaction for identity discrimination (F(3, 24)=3.52, p<.05) (Supplementary Tables 16-17 for the model and ROI comparisons).

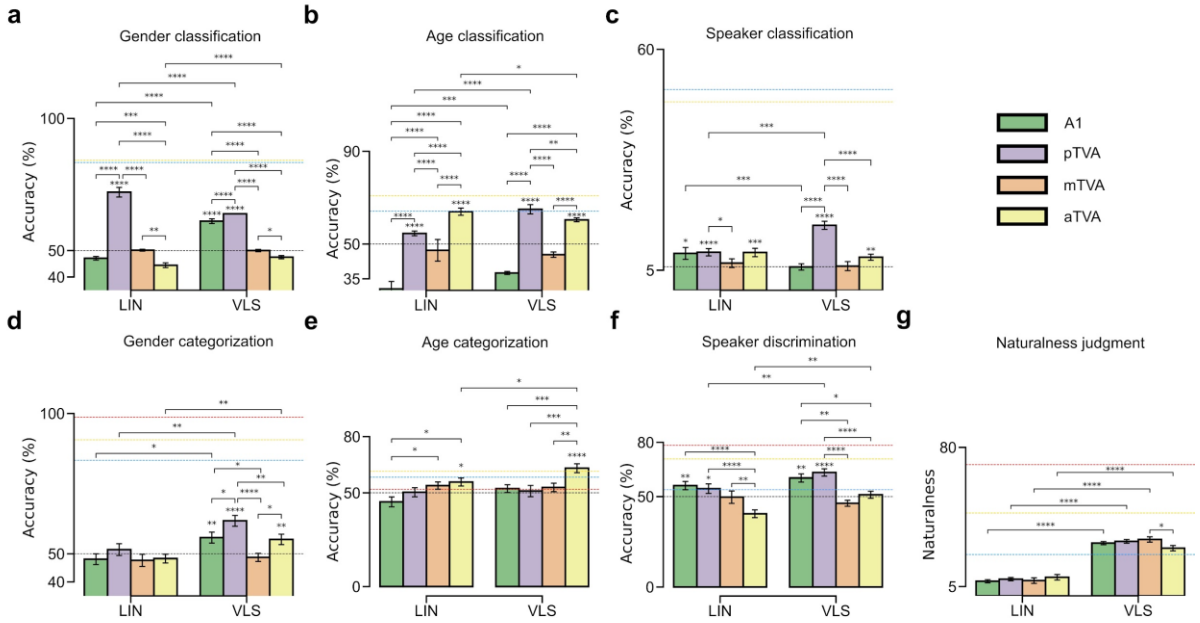


Figure 3.5: Behavioural and machine classification of the reconstructed stimuli. a,b,c, Decoding voice identity information in brain-reconstructed spectrograms. Performance of linear classifiers at categorizing speaker gender (chance level: 50%), age (chance level: 50%), and identity (17 identities, chance level: 5.88%). Error bars indicate s.e.m across 40 random classifier initializations per ROI (instance of classifiers; 2 hemispheres x 20 seeds). The horizontal black dashed line indicates the chance level. The blue and yellow dashed lines indicate the LIN and VLS ceiling levels, respectively. *p < .05; **p < .001, ***p < .001; ****p < .0001. **d,e,f,** Listener performance at categorizing speaker gender (chance level: 50%) and age (chance level: 50%), and at identity discrimination (chance level: 50%) in the brain-reconstructed stimuli. Error bars indicate s.e.m across participant scores. The horizontal black dashed line indicates the chance level, while the red, blue, and yellow dashed lines indicate the ceiling levels for the original stimuli, the LIN-reconstructed and the VLS-reconstructed, respectively. *p < .05; **p < .01; ***p < .001, ***p < .0001. **g,** Perceptual ratings of voice naturalness in the brain-reconstructed stimuli' as assessed by human listeners. *p < .05, ****p < .0001.

4. Methods

4.1. Experimental procedure overview

Three participants attended 13 MRI sessions each. The first session was dedicated to acquiring high-resolution structural data and identifying each participant's voice-selective areas using a 'voice localizer' based on different stimuli than those in the same experiment (Pernet et al., 2015; see below).

The next 12 sessions began with the acquisition of two fast structural scans for inter-session realignment purposes, followed by six functional runs, during which the main stimulus set of the experiment was presented. Each functional run lasted approximately 12 minutes. Participants 1 and 2 attended all scanning sessions (72 functional runs in total); due to technical issues, Participant 3 only performed 24 runs.

Participants were instructed to stay in the scanner while listening to the stimuli. To maintain participants' awareness during functional scanning, they were asked to press an MRI-compatible button each time they heard the same stimulus two times in a row, a rare event occurring 3% of the time (correct button hits (median accuracy ± s.d.): S1=96.67±7.10, S2=100.00±0.89, S3=95.00±3.68).

Scanning sessions were spaced by at least two days to avoid possible auditory fatigue due to the exposure to scanner noise. To ensure that participants' hearing abilities did not vary across scanning sessions, hearing thresholds were measured before each session using a standard audiometric procedure (Martin & Champlin, 2000; ISO, 2004) and compared with the thresholds obtained prior to the first session.

4.2. Participants

This study was part of the project 'Réseaux du Langage' and was promoted by the National Center for Scientific Research (CNRS). It was given approval by the local ethics committee (Comité de Protection des Personnes Sud-Méditerranée) on 13th February 2019. The National Agency for Medicines (ANSM) has been informed of this study, registered under 2017-A03614-49. Three native French human speakers (all females, 26-33 years old) were scanned. Participants gave written informed consent and received a compensation of 40€ per hour for their participation. All were right-handed, and no one had a hearing disorder or neurological disease. All participants had normal hearing thresholds of 15 dB HL for octave frequencies between 0.125 and 8 kHz.

4.3. Stimuli

The auditory stimuli were divided into two sequences. One 'voice localizer' sequence to identify the voice-selective areas of each participant (Pernet et al., 2015) and main voice stimuli.

Voice localizer stimuli. The voice localizer stimuli consisted of 96 complex sounds of 500 ms grouped into four categories: human voice, macaque vocalizations, marmoset vocalizations, and complex non-vocal sounds (more details in Bodin et al., 2021).

Main voice stimuli. The main stimulus set consisted of brief human voice sounds sampled from the Common Voice dataset (Ardila et al., 2020). Stimuli were organized into four main category levels: language (English, French, Spanish, Deutch, Polish, Portuguese, Russian, Chinese), gender (female/male), age (young/adult; young: teenagers and twenties; adult: thirties to sixties included) and identity (S1: 135 identities; S2: 142 identities; S3: 128 identities; ~44 samples per identity). Throughout the manuscript, the term 'gender' rather than 'sex' was utilized in reference to the demographic information obtained from the participants of the Common Voice dataset (Ardila et al., 2020), as it was the terminology employed in the survey ('male/female/other'). Stimulus sets differed for each participant, and the number of stimuli per set also varied slightly (number of unique stimuli: Participant 1, N=6150; Participant 2, N=6148; Participant 3, N=5123). For each participant, six stimuli were selected randomly among the sounds having high energy (as measured with the amplitude envelope) from their stimulus set and were repeated extensively (60 times) to improve the performance of the brain decoding (VanRullen et Reddy, 2019; Horikawa & Kamitani, 2017; Chang et al., 2019); these will be called the "repeated" stimuli hereafter, the remaining stimuli were presented twice. The third participant attended 5 BrainVoice sessions instead of 12, one BrainVoice session corresponding to 1030 stimuli (1024 unique stimuli and 6 'test' stimuli). Specifically, 5270 stimuli were presented to the third participant instead of ~12,000 for the two others. Among these 5270 stimuli, 5120 unique stimuli were presented once, and for the two other participants, 6 'test' stimuli were presented 25 times (150 trials). The stimuli were balanced within each run according to language, gender, age, and identity to avoid any potential adaptation effect. In addition, identity was balanced across sessions.

All stimuli of the main set were resampled at 24414 Hz and adjusted in duration (250 ms). For each stimulus, a fade-in and a fade-out were applied with a 15 ms cosine ramp to their onset and offset and were normalized by dividing the root mean square amplitude. During fMRI sessions, stimulus presentations were controlled using custom Matlab scripts (Mathworks, Natick, MA, USA) interfaced with an RM1 Mobile Processor (Tucker-

David Technologies, Alachua, USA). The auditory stimuli were delivered pseudo-randomly through MRI-compatible earphones (S14, SensiMetrics, USA) at a comfortable sound pressure level that allowed clear and intelligible listening.

4.4. Computational models

We used two computational models to learn representational space for voice signals: linear autoencoder (LIN) and deep variational autoencoder (VAE; Kingma et al., 2014). Both are encoder-decoder models that are learned to reproduce their input at their output while going through a low dimensional representation space usually called latent space (that we will call voice latent space since they are learned on voice data). The autoencoders were trained on a dataset of 182K sounds from the Common Voice dataset (Ardila et al., 2020), balanced in gender, language, and identity to reduce the bias in the synthesis (Gutierrez et al., 2021). Both models operate on sounds, which are represented as spectrograms that we describe below. These representations were tested in all the encoding/decoding and RSA analyses.

4.5. Spectrograms

We used amplitude spectrograms as input for the models that we describe below. Short-term Fourier transforms of the waveform were computed using a sliding window of 50 ms with a hop size of 12.5 ms (hence an overlap of 37.5 ms) and applying a Hamming window of size 800 samples before computing the Fourier transform of each slice. Only the magnitude of the spectrogram was kept, and the phase of the complex representation was removed. In the end, a 250 ms sound is represented by a 21×401 matrix with 21-time steps and 401 frequency bins.

We used a custom code based on $numpy.fft$ package (Harris et al., 2020). The size and the overlap between the sliding windows of the spectrogram were chosen to conform with the uncertainty principle between time and frequency resolution. The main constraint was finding a trade-off between accurate phase reconstruction with the Griffin & Lim algorithm (1983) and a reasonable spectrogram size.

We standardized each of the 401 frequency bands separately by centering all the data corresponding to each frequency band at every time step in all spectrograms, which involved removing their mean and dividing by their standard deviation. This separate

standardization of frequency bands resulted in a smaller reconstruction error compared to standardizing across all the bands.

4.6. Deep neural network

We designed a deep variational autoencoder (VAE; Kingma & Welling, 2014) of 15 layers with an intermediate hidden representation of 128 neurons that we refer to as the *voice latent space* (VLS). In an autoencoder model, the two sub-network components, the *Encoder* and the *Decoder*, are jointly learned on complementary tasks (Figure 3.1a). The Encoder network (noted Enc hereafter; 7 layers) learns to map an input, s (a spectrogram of a sound) onto a (128-dimensional) *voice latent space* representation (z; in blue in the middle of Figure 3.1a), while the Decoder (noted Dec hereafter; 7 layers) aims at reconstructing the spectrogram s from z. The learning objective of the full model is to make the output spectrogram $Dec(Enc(s))$ as close as possible to the original one s. This reconstruction objective is defined as the L2 loss, $||Dec(Enc(s)) - s||^2$. The parameters of the Encoder and of the Decoder are jointly learned using gradient descent to optimize the average L2 loss computed on the training set $\sum_{s \in Training\ Set} ||Dec(Enc(s)) - s||^2$. We trained this DNN on the Common Voice dataset (Ardila et al., 2020) according to VAE learning procedure (as explained in Kingma et Welling., 2019) until convergence (network architecture and particularities of the training procedure are provided in Supplementary Table 1), using the PyTorch python package (Paszke et al., 2019).

4.7. Linear autoencoder

We trained a linear autoencoder on the same dataset (described above) to serve as a linear baseline. Both the *Encoder* and the *Decoder* networks consisted of a single fully connected layer without any activation functions. Similar to the VAE, the latent space obtained from the *Encoder* was a 128-dimensional vector. The parameters of both the *Encoder* and the *Decoder* were jointly learned using gradient descent to optimize the average L2 loss computed on the training set.

4.8. Neuroimaging data acquisition

Participants were scanned using a 3 Tesla Prisma scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 64-channel receiver head-coil. Their movements were monitored during the acquisition using the software FIRMM (Dosenbach et al., 2017). The whole-head high-resolution structural scan acquired during the first session

was a T1-weighted multi-echo MPRAGE (MEMPRAGE) (TR = 2.5 s, TE = 2.53, 4.28, 6.07, 7.86 ms, TI=1000 ms flip angle: 8°, matrix size = 208 × 300 × 320; resolution 0.8 × 0.8 × 0.8 mm^3, acquisition time: 8min22s). Lower resolution scans acquired during all other sessions were T1-weighted MPRAGE scans (TR = 2.3 s, TE = 2.88 ms, TI=900ms, flip angle: 9°, matrix size = 192 × 240 × 256; resolution 1 × 1 × 1 mm^3, sparse sampling with 2.8 times undersampling and compressed sensing reconstruction, acquisition time: 2min37). Functional imaging was performed using an EPI sequence (multiband factor = 5 , TR = 462 ms, TE = 31.2 ms, flip angle: 45°, matrix size = 84 × 84 × 35, resolution 2.5 × 2.5 × 2.5 mm^3). Functional slices were oriented parallel to the lateral sulci with a z-axis coverage of 87.5 mm, allowing it to fully cover both the TVAs (Pernet et al., 2015) and the FVAs (Aglieri et al., 2018). The physiological signals (heart rate and respiration) were measured with Siemens' external sensors.

4.9. Pre-processing of neuroimaging data and general linear modeling

Tissue segmentation and brain extraction were performed on the structural scans using the default segmentation procedure of SPM 12 (Ashburner et al., 2012). The preprocessing of the BOLD responses involved correcting motion, registering inter-runs, detrending, and smoothing the data. Each functional volume was realigned to a reference volume taken from a steady period in the session that was spatially the closest to the average of all sessions. Transformation matrices between anatomical and functional data were computed using boundary-based registration (FSL; Smith et al., 2004). The data were respectively detrended and smoothed using the *nilearn* functions *clean_img* and *smooth_img* (kernel size of 3mm) (Abraham et al., 2014), resulting in the matrix $Y \in R^{S \times V}$, with S the number of scans and V the number of voxels.

A first general linear model (GLM) was fit to regress out the noise by predicting Y from a "denoised" design matrix, composed of $R = 38$ regressors of nuisance (Supplementary Figure S4). These regressors of nuisance, also called covariates of no interest, included: 6 head motion parameters (3 variables for the translations, 3 variables for the rotations); 18 'RETROICOR' regressors (Glover et al., 2000) using the *TAPAS PhysIO* package (Kasper et al., 2017) (with the hyperparameters set as specified in Snoek et al.) were computed from the physiological signals; 13 regressors modeling slow artifactual trends (sines and

cosines, cut frequency of the high-pass filter = 0.01 Hz); and a confound-mean predictor. The design matrix was convolved with a hemodynamic response function (HRF) with a peak at 6 s and an undershoot at 16 s (Glover et al., 1999); we note the convolved design matrix as $X_d \in R^{S \times R}$. The "denoise" GLM's parameters $\beta_d \in R^{R \times V}$ were optimized to minimize the amplitude of the residual $\beta_d = argmin_{\beta \in R^{R \times V}} \; || \; Y \; - X_d \, \beta \; ||^2$. We used a lag-1 autoregressive model (ar(1)) to model the temporal structure of the noise (Friston et al., 2002). The *denoised* BOLD signal Y_d was then obtained from the original one according to $Y_d = Y - (X_d \, \beta_d) \in R^{S \times V}$.

A second "stimulus" GLM model was used to predict the denoised BOLD responses for each stimulus using a design matrix $X_s \in R^{S \times (N_S + 1)}$ (which was convolved with an hemodynamic response function, HRF as above) and a parameters matrix $\beta_s \in R^{(N_S + 1) \times V}$ where N_S stands for the number of stimuli. The last row (resp. column) of β_s (resp. X_s) stands for a silence condition. Again, β_s was learned to minimize the residual $\beta_s = argmin_{\beta \in R^{(N_S+1) \times V}} \; || \, Y_d - X_s \, \beta \; ||^2$. Once learned, each of the first N_s line of β_s was corrected by subtracting the $(N_s+1)^{th}$ line, yielding the contrast maps for stimuli $\tilde{\beta}_s \in R^{N_S \times V}$. We note hereafter $\tilde{\beta}_s[i, :] \in R^V$ the contrast map for a given stimulus, it is the i^{th} line of $\tilde{\beta}_s$.

A third "identity" GLM was fit to predict the BOLD responses of each voice speaker identity, using a design matrix $\beta_i \in R^{(N_i + 1) \times V}$ and a design matrix $X_i \in R^{S \times (N_i + 1)}$ (which was again convolved with an hemodynamic response function, HRF) where N_s stands for the number of unique speakers. Again the last row/column in β_i and X_i stands for the silent condition. β_i is learned to minimize the residual $\beta_i = argmin_{\beta \in R^{(N_i+1) \times V}} ||Y_d - X_i \, \beta \; ||^2$ (Supplementary Figure S3a). Again, the final speaker contrast maps were obtained by contrasting (i.e., subtracting) the regression coefficients in a row of β_i with the silence condition (last row; Supplementary Figure S3a), yielding $\tilde{\beta}_i \in R^{N_s \times V}$. Here the j^{th} row of $\tilde{\beta}_i$, $\tilde{\beta}_i[j, :] \in R^V$, represents the amplitude of the BOLD response of the contrast map for speaker j (i.e., to all the stimuli from this speaker).

A fourth "localizer" GLM model was used to predict the denoised BOLD responses of each sound category from the *Voice localizer stimuli* presented above. The procedure was

similar to that described for the two previous GLM models. Once the GLM was learned, we contrasted the human voice category with the other sound categories in order to localize for each participant the posterior Temporal Voice Area (pTVA), medial Temporal Voice Area (mTVA), and anterior Temporal Voice Area (aTVA) in each hemisphere. The center of each TVA corresponded to the local maximum of the voice > nonvoice t-map whose coordinates were the closest to the TVAs reported (Pernet et al., 2015). The analyses were carried out for each region of interest (ROI) of each hemisphere.

Additionally, we defined for each participant the primary auditory cortex (A1) as the maximum value of the probabilistic map (non-linearly registered to each participant functional space) of Heschl's gyri provided with the MNI152 template (Penhune et al., 1996), intersected with the sound vs silence contrast map.

4.10. Identity-based and stimulus-based representations

We performed analyses either at the stimulus level, e.g., predicting the neural activity of a participant listening to a given *stimulus* ($\tilde{\beta}_s$'s lines) from the *voice latent space* representation of this stimuli, or at the speaker identity level, e.g., predicting the average neural activity in response to stimuli of a given speaker *identity* ($\tilde{\beta}_i$'s lines) from this speaker's *voice latent space* representation. The identity-based analyses were used for the characterization of the *voice latent space* (Figure 3.1), the brain encoding (Figure 3.2), and the representational similarity analysis (Figure 3.3), while the stimulus-based analyses were used for the brain decoding analyses (Figure 3.4, 5).

We conducted stimulus-based analyses to examine the relationship between stimulus contrast maps in neural activity ($\tilde{\beta}_s$) and the encodings of individual stimulus spectrograms computed by the encoder of an autoencoder model (either linear or deep variational autoencoder) on the computational side. We will note ${z_s}^{lin} \in R^{N_s \times 128}$ encodings of stimuli by the LIN model and ${z_s}^{vae} \in R^{N_s \times 128}$ the encodings of stimuli computed by the VAE model. The encoding of the kth stimuli by one of these models is the kth row of the corresponding matrix, and it is noted as ${z_s}^{model}[k, :]$.

For identity-based analyses, we studied relationships between identity contrast maps in $\tilde{\beta}_i$ on the neural activity side and an encoding of speaker identity in the VLS implemented by an autoencoder model (LIN or VAE) on the computational side, e.g., we note ${z_i}^{vae}[j]$ the representation of speaker *j* as computed by the *vae* model. We chose to define a speaker identity-based representation as the average of a set of sample-based representations for stimuli from this speaker, e.g., ${z_i}^{model}[j] = 1/|S_j| \sum_{k \in S_j} {z_s}^{model}[k,:]$ where S_j stands for the set of stimuli by speaker *j* and *model* stands for *vae* or *lin*. Averaging in the *voice latent space* is expected to be much more powerful and relevant than averaging in the input space spectrograms (VanRullen & Reddy, 2019).

4.11. Characterization of the autoencoder latent space

We characterized the organization of the *voice latent space* (VLS) and of the features computed by the linear autoencoder (LIN) by measuring through classification experiments the presence of information about the speaker's gender, age, and identity in the representations learned by these models.

We first computed the speaker's identity *voice latent space* representations for each of the 405 speakers in the main voice dataset (135+142+128 see *Stimuli* section) as explained above.

Next, we used these speakers' voice latent space representation to investigate if gender, age, and identity were encoded in the VLS. To do so, we divided the data into separate train and test sets and learned classifiers to predict gender, age, or identity from the train set. The balanced (to avoid the small effects associated with unbalanced folds) accuracy of the classifiers was then evaluated on the test set. The higher the performance on the test set, the more confident we are that the information is encoded in the VLS. More specifically, for each task (gender, age, identity), we trained a Logistic Regression classifier (linear regularized logistic regression; L2 penalty, tol=0.0001, fit_intercept=True, intercept_scaling=1, max_iter=100) using the scikit-learn python package (Pedregosa et al., 2018).

In order to statistically evaluate the significance of the results and to avoid potential overfitting, the classifications were repeated 20 times with 20 different initializations (seed), and the metrics were then averaged for each voice category (gender, age). More specifically, we repeated the following experiment 20 times with 20 different random seeds. For each seed, we performed 5 train-test splits, with 80% of the data in the training and 20% in the test set. For each split, we used 5-fold cross-validation on the training set to select the optimal value for the regularization hyperparameter C (searching between 10 values logarithmically spaced on the interval [-3, +3]). We then computed the generalization performance on the test set of the model trained on the full training set with the best hyperparameter value. Reported results were then averaged over 20 experiments. Note that data were systematically normalized with a scaler fitted on the training set. We used a robust scaling strategy for these experiments (removing the median, then scaling to the quantile range, 25th quantile, and 75th quantile), which occurs to be more relevant with a small training set.

To investigate how speaker identity information is encoded in the latent space representations of speakers' voices, we computed speaker identity *voice latent space* representations by averaging 20 stimulus-based representations in order to obtain a limited amount of data per identity that could be distributed across training and test datasets.

We first tested whether the mean of the distribution of accuracy scores obtained for 20 seeds was significantly above the chance level using one-sample t-tests. We then evaluated the difference in classification accuracy between the VLS and LIN via one-way ANOVAs (dependent variable: test balance accuracy; between factor: Feature) for each category (speaker gender, age, identity). We performed post-hoc planned paired t-tests between the models to test the significance of the VLS-LIN difference.

4.12. Brain encoding

We performed encoding experiments on identity-based representations for each of the three participants (Figure 3.2). For each participant, we explored the ability to learn a regularized linear regression that predicts a speaker-based neural activity, e.g. the j^{th} speaker's contrast map $\tilde{\beta}_i[j] \in R^V$, from this speaker's voice latent space representation,

that we note ${z_i}^{model}[j] \in R^{128}$ (Figure 3.2a). We carried out these regression analyses for each ROI (A1, pTVA, mTVA, aTVA) in each hemisphere and participant, independently.

The regression model parameters $\widehat{W}_{encod} \in R^{128 \times V}$ were learned according to the following:

$$\widehat{W}_{encod} = argmin_{W_{encod} \in R^{128 \times V}} \sum_{j=1..N_i} ({z_i}^{model}[j] \times W_{encod} - \tilde{\beta}_i[j])^2 + \lambda \|W_{encod}\|^2$$

where λ is a hyperparameter tuning the optimal tradeoff between the data fit and the penalization terms above. We used the ridge regression with built-in cross-validation as implemented as $RidgeCV$ in the scikit-learn library (Pedregosa et al., 2018).

The statistical significance of each result was assessed with the following procedure. We repeated the following experiment 20 times with different random seeds. Each time, we performed 5 train-test splits, with 80% of the data in the training and 20% in the test set. For each split, we used RidgeCV (relying on leave-one-out) on the training set to select the optimal value for the hyperparameter λ (searching between 10 values logarithmically spaced on the interval $[10^{-1}; 10^{8}]$). Following standard practice in machine learning, we then computed the generalization performance on the test set of the model trained on the full training set with the best hyperparameter value. Reported results are then averaged over 20 experiments. Note that here again, with small training sets, data were systematically normalized in each experiment using robust scaling.

The evaluation relied on the 'brain score' procedure (Schrimpf et al., 2018), which evaluates the performance of the ridge regression with a Pearson's correlation score. Correlations between measured neural activities $\tilde{\beta}_i$ and predicted ones $\widehat{{z_i}^{model}} * W_{encod}$ were computed for each voxel and averaged over repeated experiments (folds and seeds), yielding one correlation value for every voxel and for every setting. The significance of the results was assessed with one-sample t-tests for the Fisher z-transformed correlation scores (3 x participants x 2 hemispheres x V voxels). For each region of interest, the scores are reported across participants and hemispheres (Figure 3.2b). The exact same procedure was followed for the LIN modeling.

In order to determine which of the two feature spaces (VLS, LIN) and which of the two ROI (A1, TVAs) yielded the best prediction of neural activity, we compared the means of distributions of correlations coefficients using a mixed ANOVA performed on the Fisher z-transformed coefficients (dependent variable: correlation; between factor: ROI; repeated measurements: Feature; between-participant identifier: voxel).

For each ROI, we then used t-tests to perform post-hoc contrasts for the VLS-LIN difference in brain encoding performance (comparison tests in Figure 3.2b; Supplementary Table 4). We finally conducted two-sample t-tests between the brain encoding model's scores trained to predict A1 and those trained to predict temporal voice areas to test the significance of the A1-TVAs difference (Supplementary Table 5).

The statistical tests were all performed using the *pingouin* python package (Vallat., 2018).

4.13. Representational similarity analysis

The RSA analyses were carried out using the package *rsatoolbox* (Schütt et al., 2021; https://github.com/rsagroup/rsatoolbox). For each participant, region of interest, and hemisphere, we computed the cerebral Representational Dissimilarity Matrix (RDM) using Pearson's correlation between the speaker identity-specific response patterns of the GLM estimates $\tilde{\beta}_i$ (Walther et al., 2016) (Figure 3.3a). The model RDMs were built using cosine distance (Xing et al., 2015; Bhattacharya et al., 2017; Wang et al., 2018), capturing speaker pairwise feature differences predicted by the computational models LIN and the VLS (Figure 3.3a). The GLM estimates, and the computational models' features were first normalized using robust scaling for greater comparability with the rest of the analyses described here. We computed the Spearman correlation coefficients between the brain RDMs for each ROI and the two model's RDMs (Figure 3.3b). We assessed the significance of these brain-model correlation coefficients within a permutation-based 'maximum statistics' framework for multiple comparison correction (one-tailed inference; N permutations = 10,000 for each test; permutation of rows and columns of distance matrices, see Giordano et al., 2023 and Maris & Oostenveld, 2007; see Figure 3.3b). We evaluated the VLS-LIN difference using a two-way repeated-measures ANOVA on the Fisher z-transformed Spearman correlation coefficients (dependent variable:

correlation; within factors: ROI and Feature; participant identifier: participant hemisphere pair). The same permutation framework was also used to assess the significance of the difference between the RSA correlation for the VLS and LIN models.

4.14. Brain decoding

Brain decoding was investigated at the stimulus level. The stimuli's voice latent space representations $z_s^{model} \in R^{N \times 128}$ and voice samples' contrast maps $\tilde{\beta}_s \in R^{N \times V}$ were divided into train and test splits, normalized across voice samples using robust scaling, then fit to the training set. For every participant and each ROI, we trained a L_2-regularized linear model $W \in R^{V \times 128}$ model to predict the voice samples' latent vectors from the voice samples' contrast maps (Figure 3.4a). The hyperparameter selection and optimization were done similarly to the brain encoding scheme. Training was performed on non-repeated stimuli (see *Stimuli* section). We then used the trained models to predict for each participant the 6 repeated stimuli that were the most presented. Waveforms were estimated starting from the reconstructed spectrograms using the Griffin-Lim phase reconstruction algorithm (Griffin & Lim, 1983).

We then used classifier analyses to assess the presence of voice information (gender, age, speaker identity) in the reconstructed latent representations (i.e., the latent representation predicted from the brain activity of a participant listening to a specific stimulus) (Figure 3.5a, b, c). To this purpose, we first trained linear classifiers to categorize the training voice stimuli (participant 1, N = 6144; participant 2, N = 6142; participant 3, N = 5117; total, N = 17403) by gender (2 classes), age (2 classes) or identity (17 classes) based on VLS coordinates. Secondly, we used the previously trained classifiers to predict the identity information based on the VLS derived from the brain responses of the 18 Test voice stimuli (3 participants x 6 stimuli). We first tested using one-sample t-tests that the mean of the distribution of accuracy scores obtained across random classifier initializations of classifiers (2 hemispheres x 20 seeds = 40) was significantly above the chance level for each category, ROI, and model. We then evaluated the difference in performance at preserving identity-related information depending on the model or ROI via two-way ANOVAs (dependent variable: accuracy; between factors: Feature and ROI). We performed post-hoc planned paired t-tests between each model pair

to test the significance of the VLS-LIN difference. Two-sample t-tests were finally used to test the significance of the A1-TVAs difference.

4.15. Listening tests

We recruited 13 participants through the online platform Prolific (www.prolific.co) for a series of online behavioral experiments. All participants reported having normal hearing. These experiments aimed to evaluate how well voice identity information and naturalness are preserved in fMRI-based reconstructed voice excerpts. In the main session, participants carried out 4 tasks, in the following order: 'speaker discrimination' (~120 min), 'perceived naturalness' (~30 min), 'gender categorization' (~30 min), 'age categorization' (~30 min). The experiment lasted 3 hours and 35 minutes, and each participant was paid £48.

Prior to the main experiment session, participants carried out a short loudness-change detection task to ensure they wore headphones, were attentive, and were correctly set up for the main experiment (Woods et al., 2017). On each of the 12 trials, participants heard 3 tones and were asked to identify which tone was the least loud by clicking one of 3 response buttons: 'First', 'Second', or 'Third'. Participants were admitted to the main experiment only if they achieved perfect performance in this task. We refined the participant pool by excluding those who performed poorly on the original stimuli.

The following three tasks were each carried out on the same set of 342 experimental stimuli, each presented on a different trial: 18 original stimuli, 36 stimuli reconstructed directly from the LIN and the VLS models, and 18 stimuli x 2 models x 4 regions of interest x 2 hemispheres= 288 brain-reconstructed stimuli. In the 'perceived naturalness' task, participants were asked to rate how natural the voice sounded on a scale ranging from 'Not at all natural' to 'Highly natural' (i.e., similar to a real recording) and were instructed to use the full range of the scale. During the 'gender categorization' task, participants categorized the gender by clicking on a 'Female' or 'Male' button. Finally, in the 'age categorization' task, participants categorized the speaker's age by clicking on a 'Younger' or 'Older' button. In the 'speaker discrimination' task, participants carried out 684 trials (342 experimental stimuli x 2) with short breaks in between. In each trial, they were

presented with 2 short sound stimuli, one after the other, and participants had to indicate whether they were from the same speaker.

To evaluate the participants' performance, we first conducted one-sample t-tests to examine whether the mean accuracy score calculated from their responses was significantly higher than the chance level for each model and ROI. Next, we used two-way repeated-measures ANOVAs to assess the variation in participants' performances in identifying identity-related information (dependent variable: accuracy; between-participant factors: Feature and ROI). To determine the statistical significance of the VLS-LIN difference, we carried out post-hoc planned paired t-tests between each model pair. Finally, we employed two-sample t-tests to evaluate the statistical significance of the A1-TVAs difference.

5. Conclusion

In this Chapter, we examined to what extent the cerebral activity elicited by brief voice stimuli can be explained by machine-learned representational spaces, specifically focusing on identity-related information. We trained a linear model and a DNN model to reconstruct 100,000s of short voice samples from 100+ speakers, providing low-dimensional spaces (LIN and VLS), which we related to fMRI measures of cerebral response to thousands of these stimuli. We find: (i) that 128 dimensions are sufficient to explain a sizeable portion of the brain activity elicited by the voice samples and yield brain-based voice reconstructions that preserve identity-related information; (ii) that the DNN-derived VLS outperforms the LIN space, particularly in yielding more brain-like representational spaces and more naturalistic voice reconstructions; (iii) that different ROIs have different degrees of brain-model relationship, with marked differences between A1 and the a, m, and pTVAs.

Low-dimensional spaces generated by machine learning have been used to approximate cerebral face representations and reconstruct recognizable faces based on fMRI (VanRullen et Reddy, 2019; Dado et al., 2022). In the auditory domain, however, they have mainly been used with a focus on linguistic (speech) information, ignoring identity-related information (but see Akbari et al., 2019). Here, we applied them to brief voice stimuli–with minimal linguistic content but already rich identity-related information–and

found that as little as 128 dimensions account reasonably well for the complexity of cerebral responses to thousands of these voice samples as measured by fMRI (Figure 3.2). LIN and VLS both showed brain-like representational geometries, particularly the VLS in the aTVAs (Figure 3.3). They made possible what is, to our knowledge, the first fMRI-based voice reconstructions to preserve voice-related identity information such as gender, age, or even individual identity, as indicated by above-chance categorization or discrimination performance by both machine classifiers (Figure 3.5a-c) and human listeners (Figure 3.5d-f). Note that LIN and VLS also represent the limited linguistic content of the brief stimuli, as indicated by high language classification performance (Supplementary Figure S2).

Estimation of fMRI responses (encoding) by LIN yielded correlations largely comparable to those by VLS (Figure 3.2b), although many voxels were only explained by one or the other space (Figure 3.2c). However, in the RSA, VLS yielded higher overall correlations with brain RDMs (Figure 3.3), suggesting a representational geometry closer to that instantiated in the brain than LIN. Further, VLS-reconstructed stimuli sounded more natural than the LIN-reconstructed ones (Figure 3.5g) and yielded both the best speaker discrimination by listeners (Figure 3.5f) and speaker classification by machine classifiers (Figure 3.5c). Unlike LIN, which was generated via linear transforms, VLS was obtained through a series of nonlinear transformations (Wetzel, 2017). The fact that the VLS outperforms LIN in decoding performance indicates that nonlinear transformation is required to better account for the brain representation of voices (Naselaris et al., 2011; Cowen et al., 2014; Han et al., 2019).

Comparisons between ROIs revealed important differences between A1 and the a, m, and pTVAs. For both LIN and VLS, fMRI signal (encoding) predictions were more accurate for the mTVAs than for A1, and for A1 than for the pTVAs (Figure 3.2b). The aTVAs yielded the highest correlations with the models in the RSA (Figure 3.3). Stimulus reconstructions (Figure 3.4) based on the TVAs also yielded better gender, age, and identity classification than those based on A1, with gender and identity best preserved in the pTVA-, and to a lesser extent, in the aTVA-based reconstructions (Figure 3.5). These results show that the a and pTVAs not only respond more strongly to vocal sounds than A1, but they also represent identity-related information in voice better than mTVA, which was previously

anticipated in some neuroimaging studies (Latinus et al., 2011; Charest et al., 2013; Aglieri et al., 2021).

Overall, this chapter shows that a DNN-derived representational space provides an interesting approximation of the cerebral representations of brief voice stimuli that can preserve identity-related information. We find it remarkable that such results could be obtained to explain sound representations despite the poor temporal resolution of fMRI. Future work combining more complex architectures to time-resolved measures of cerebral activity, such as magneto-encephalography (Défossez et al., 2023) or ECoG (Pasley et al., 2012), will likely yield better models of the cerebral representations of voice information.

Discussion

In this thesis, we employed artificial intelligence (AI) in two distinct yet complementary ways: AI as a computational model and AI as a tool.

In Chapter 1, we proposed a synthesized model for human voice processing. We reviewed older and recent studies on human voice processing, suggesting a potential role for each voice patch within each hemisphere. Based on this model, we explored the voice patch system across primate brains, including humans, macaques, and marmosets.

In Chapter 2, we created a large-scale dataset of marmoset vocalizations. We employed AI as a tool to detect, segment, and label marmoset vocalizations from raw recordings. The dataset and the trained classifier will be publicly available for future research in vocal communication.

In Chapter 3, we employed AI as a computational model to demonstrate that voice representations derived from deep neural networks constitute interesting approximations of cerebral representations and can significantly predict brain activity in response to voice, as recorded with fMRI. Additionally, we reconstructed the spectrograms of stimuli presented to the participants. We retrieved voice identity information from the reconstructed stimuli using machine learning classifiers and behavioral tasks performed by human listeners.

Overall, our findings underscore the potential of AI in shedding light on the brain's voice processing mechanisms, serving both as a computational model and as a tool.

1. Evolutionary origins of voice perception

What is the functional role of each unit within the "voice patch" system in the primate brain when processing vocal information?

In Chapter 1, we reviewed older and recent literature on voice processing in humans and non-human primates to determine the potential role of each voice-sensitive area. We proposed a synthesized voice processing model based on brain studies in primates that outlines a pathway with three stages: detection, measurement, and categorization for voice recognition. The model tentatively underscores the roles of the fronto-temporal-limbic network and the hemispheric specialization, where the right predominantly handles voice identity, the left manages semantic deciphering, and the limbic system, the vocal emotion, bilaterally. Differentiating computational phases—detection, measurement, and categorization—offer a granular understanding of voice perception.

However, I have identified several essential questions in Chapter 1 that need to be answered, which I will discuss in the following subsections.

1.1. Investigating voice cell coding

Standard methods for exploring the neural mechanisms of voice processing can identify broad neural substrates but often offer limited insight into the overlapping, segregation, and form of neuronal representations involved in processes like identity recognition (Perrodin et al., 2015). This limitation arises because neuroimaging techniques typically measure either surrogate markers of neuronal activity or large-scale neural responses (such as fMRI; see *Functional neuroimaging* in the Introduction). Therefore, there is a need for direct measures of localized neuronal computations. Direct neuronal recordings, such as depth electrode recordings or electrocorticography, in human patients undergoing monitoring for neurosurgery brought critical insights into neuronal functions within localized auditory regions of the human brain (Zhang et al., 2021; Rupp et al., 2022). Meanwhile, research in animal models allows for the examination of neuronal processes at multiple scales directly within the regions of interest and provides more specificity in neuronal manipulation (activation and/or inactivation) (fMRI-guided electrophysiology, Perrodin et al., 2011; Giamundo et al., *submitted*). However, the advancements in animal research have not kept pace (Perrodin et al., 2015).

Within the ventral visual stream of humans and non-human primates, faces seem to be represented within a system known as the face patch system, which shares many

similarities with the voice patch system (Belin, 2017). However, the visual domain has benefited from a larger number of fMRI-guided studies investigating the neural codes of the face patch system. Within these patches, faces are represented using low-dimensional neural codes, where each neuron encodes an orthogonal axis of variation in face space (Chang & Tsao, 2017). A recent study built upon this finding using self-supervised generative models (Higgins et al., 2021). Their model succeeded in "disentangling" face images into meaningful factors of variation (e.g., hair length, gender) and establishing a one-to-one correspondence between these factors and the responses of face single units. A significant avenue for future research in brain-based voice processing would be to explore the neural code of the voice patch system, especially to determine if it employs similar low-dimensional codes as observed for faces.

1.2. Mapping voice patch connections and processing stages

How do voice patches connect within the brain, what are their processing stages, and are there distinct temporal dynamics in voice processing? Various studies have provided insights into the temporal dynamics of voice processing in the temporal voice areas, predominantly in humans (Charest et al., 2009; Capilla et al., 2013; Schall et al., 2015; Zhang et al., 2021; Norman-Haignere et al., 2022; Rupp et al., 2022). In our synthesized model, we highlighted a key finding: voice processing initiates bilaterally in the mid-temporal voice areas and then proceeds to both the posterior and anterior voice areas in parallel (Figure 1.1). However, only a handful of studies have shed light on the dynamics of the frontal voice areas (Lowe et al., 2021) and the vocal emotion responses in the limbic system (Giordano et al., 2021). Further research targeting these areas is required to enhance our understanding of their temporal dynamics in voice processing.

1.3. Employing DNNs as a computational model

How accurately do DNNs approximate human processes, especially in the temporal and frontal voice areas? With DNNs emerging as promising models for voice processing, we suggest in Chapter 3 to employ DNNs to computationally simulate (1) the processes of voice detection, measurement, and categorization; (2) the presumed functional roles of the voice areas, especially focusing on the motor/semantic, voice identity, and vocal emotion processing axes. By training DNNs on extensive voice datasets, one could mimic

the hierarchical processing of voice information, ranging from basic detection to high-level categorization.

Can combining bio-inspired DNNs with hierarchical voice models provide deeper insights into primate voice processing? Another potential pathway is to examine various architectures and biases to discern what contributes to the similarities between the brain and DNNs. One straightforward method might be to pre-map the cortical areas in the model's architecture (Kubilius, Schrimpf, et al., 2018) or to train the model to emulate the brain's statistical properties (Cadena et al., 2019; Federer et al., 2020), or even to amalgamate these strategies with an additional training objective that aligns the pre-mapped cortical areas with their corresponding brain responses. Although these methods have been explored in the visual domain, they have yet to be applied in the auditory domain.

2. Encoding and decoding of voice identity

Would DNNs provide reasonable approximations of cerebral representations? — in particular regarding the processing of voice identity. In Chapter 3, we tried to address this question by applying representation learning to model the voice signal in a representation that might correlate with the brain activity associated with different voice stimuli. In particular, we tried to answer ***where this representation exhibits the strongest alignment with actual cerebral voice-related activity within the auditory cortex.*** We found that the voice latent space (VLS) derived from deep neural networks captures key aspects of voice stimuli and performs better than traditional linear methods like PCA in decoding voice-related brain activity. This indicates the advantage of using nonlinear models to understand brain representations of voices. Moreover, the relationship between VLS and TVAs was stronger than that with A1, highlighting the importance of TVAs in voice identity processing. There was also evidence of right-hemispheric lateralization for speaker identity processing. Our findings demonstrate that deep learning-derived representations provide an effective representation of voice identity information in the voice-selective areas of the auditory cortex.

However, future research in the field would benefit from several key improvements, which I will discuss in the following subsections.

2.1. Incorporating the extended voice system

In Chapter 3 and the preceding section, we investigated the most famous voice-sensitive areas within the temporal lobe (TVAs). However, as emphasized in Chapter 1, the extended voice system plays a crucial role in processing higher-level information regarding voice identity (within the frontal lobe, FVAs) and voice emotions (within the limbic system). For a comprehensive understanding of the post-perceptual processing of high-level voice information, future research should include these regions of interest in spatially and temporally specific experiments.

2.2. Developing better models and evaluations

Even with the recent progress in building more powerful models using DNNs, brain activity prediction is not perfect, with correlation scores far from 1. This could be partly because of the intrinsic noise in functional neuroimaging data (fMRI, MEG), but a big part of the variability in brain activity still is not explained. In this section, we introduce different ways to address this challenge.

2.2.1. Learning better representations

In our work, we employed a two-stage procedure to align our computational model with brain activity: (i) initially, we utilized representation learning, specifically unsupervised learning, by training autoencoder models to compress and then reconstruct voice spectrograms; (ii) subsequently, we employed a regularized linear regression model to map the autoencoder's representations of the stimuli with the brain responses to the same stimuli. While this methodology underscored the similarities between voice representation and brain responses, newer techniques present interesting avenues to explore, mainly in the visual domain literature.

Chen et al. (2023) suggested initially learning a self-supervised representation of fMRI data using mask modeling in a large latent space. By enhancing a latent diffusion model with conditioning, they demonstrated that their model could reconstruct highly plausible images with semantically matching details from brain recordings, using only a few paired annotations.

Alternatively, replacing the regression objective with a contrastive objective has also shown promise in brain decoding (Scotti, Banerjee, et al., 2023, *preprint*; Défossez et al., *preprint*). Lee, Lee, et al. (2023, *preprint*) recently utilized EEG and a generative model to decode imagined speech using EEG. In these three approaches, a key step was to train their model or part of their model directly with the brain responses, contrasting with our two-stage approach.

2.2.2. Improving model interpretability

Even with better similarity scores between the model and brain responses, the question remains: ***What insights do we gain regarding brain function?*** One possible approach is to interpret the learned model. However, the challenge of making sense of DNNs, especially when applied to complex stimuli, is significant. In our work, we were able to identify certain properties within the autoencoder's voice latent space, e.g., a major organizational dimension of the latent space is along voice gender (Figure 3.1b). However, a representation is only considered interpretable when separated into subcomponents, with each subcomponent originating from independent factors (Higgins et al., 2017) and corresponding to a real-world concept without containing information related to the others. For example, one dimension could encode gender, another the speaker's identity, age, etc. Higgins et al. (2021) leveraged disentangled representations through self-supervised generative models. The model they developed effectively "disentangled" face images into meaningful factors of variation, such as hair length and gender, and established a direct correspondence between these factors and the responses of face single units. This approach would benefit future research to develop interpretable computational models for cerebral voice processing.

2.2.3. Leveraging better brain recording techniques

Many studies currently depend on brain recording techniques that possess high spatial resolution with compromised temporal resolution (e.g., fMRI) or vice versa (e.g., M/EEG). As noted in Chapter 1, the most impactful studies highlighted in our literature review take advantage of intracranial recordings, which offer both high spatial and temporal resolutions (e.g., Zhang et al., 2021; Rupp et al., 2022). A recent advancement in non-invasive brain recording techniques is the development of optically pumped magnetometers (OPM)-MEG, a novel type of MEG equipment that offers multiple benefits

over traditional scanners (Brookes et al., 2022). These advantages include enhanced signal sensitivity, improved spatial resolution, and the freedom for participants to move during scanning. The OPM-MEG represents a significant stride towards overcoming the limitations of existing brain recording techniques and holds promise for garnering more precise insights into brain activity and its correlation with various cognitive functions.

3. Computational neuroethology of vocal communication

Chapter 2 looked at the use of computational methods to better understand the evolution of vocal communication. A remaining question from our review (Chapter 1) is: ***Do voice patches across the primate species share similar coding principles?*** To explore this, we suggest using the methodologies outlined in Chapter 3 — specifically, applying deep learning-based vocal representations and correlating them with brain responses to vocal stimuli in non-human primates. This task requires training a deep neural network with a large set of vocal signals specific to the species. Currently, such datasets for both macaque and marmoset monkeys are unavailable in the literature.

To address this gap, we presented a complete pipeline for extracting and analyzing vocalizations from marmoset monkey recordings, continuously recorded at a sampling rate of 96 kHz from a room housing about 20 marmosets in three cages. The dataset includes over 800,000 files, totaling 253 hours of data collected over 40 months. Each recording lasts a few seconds and captures the marmosets' social vocalizations, covering their entire known vocal repertoire during the experimental period. Around 215,000 calls were annotated with the vocalization type. The provided dataset, source code, and pre-trained classifier are valuable resources for future research in this field. Moving forward, it is essential to build on this initial work by including more species-specific vocal datasets and improving computational methodologies to further our understanding of the evolution of vocal communication.

Our dataset is a valuable resource for future research, enabling investigations into the functions, contexts, and variations within the marmoset vocal repertoire. Moreover, it offers opportunities for comparative studies with other primate species, including humans, to discover shared and unique aspects of vocal communication across

evolutionary lineages. Below, I will outline some suggested research directions in marmosets, but similar investigations can be carried out with other non-human primate species. We have recorded a similar dataset of macaque vocalizations and are currently employing the same methodology to compile a large-scale dataset. This will enable us to leverage a comparative approach to examine the communication systems of different primate species, thereby enriching our understanding of the evolution of vocal communication.

3.1. Characterize the acoustical properties of the marmoset vocal repertoire

Recently, there has been increased interest in the common marmoset (Callithrix jacchus) as a neuroscientific model organism (Miller et al., 2016), leading to many attempts to study and characterize their vocal repertoire (Bezerra & Souto, 2008; Pistorio et al., 2006; Zhao et al., 2019). However, both past and recent literature show that recording setups were limited to a sampling rate of 48 kHz. Benefiting from a higher sampling rate compared to previous studies, and with hundreds of thousands of samples available instead of a few thousand, a new opportunity arises to better characterize the acoustic properties of the marmoset vocal repertoire. This includes analysis of fundamental frequency (F0), length statistics per vocalization type, frequency, number of harmonics, and characterizing the formants within the marmoset vocalization repertoire (Fukushima et al., 2015; Pistorio et al., 2006; Zhao et al., 2019; Norcross & Newman, 1993).

3.2. Investigate the conversational patterns of the marmoset vocal repertoire

Recently, there has been increased interest in the common marmoset (Callithrix jacchus) as a neuroscientific model organism (Miller et al., 2016), leading to many attempts to study and characterize their vocal repertoire (Bezerra & Souto, 2008; Pistorio et al., 2006; Zhao et al., 2019). However, both past and recent literature show that recording setups were limited to a sampling rate of 48 kHz. Benefiting from a higher sampling rate compared to previous studies, and with hundreds of thousands of samples available instead of a few thousand, a new opportunity arises to better characterize the acoustic properties of the marmoset vocal repertoire. This includes analysis of

fundamental frequency (F0), length statistics per vocalization type, frequency, number of harmonics, and characterizing the formants within the marmoset vocalization repertoire (Fukushima et al., 2015; Pistorio et al., 2006; Zhao et al., 2019; Norcross & Newman, 1993).

3.3. Study the changes in marmoset vocalizations during aging

Newborn marmosets have a distinct vocalization known as infant cry or nga, which gradually changes into adult vocalizations like phee over time. By monitoring the vocalizations of the same individual over time, we can observe how marmoset vocalizations evolve as they age, such as how an infant cry slowly transitions into a phee. This could enhance our understanding of their communication system. Such a study would necessitate isolating certain individuals and recording their call types and/or behaviors to identify their vocalizations within our dataset.

3.4. Linking vocalization with behavior

Establishing a causal link between vocal communication and behavior is a significant step toward comprehensively understanding a species' social interactions (Prat, 2019; Fischer et al., 2021). ***Can we predict an individual's behavior based on the vocalization sequences of the group or individuals?*** Several studies have explored this question in marmosets (Bezerra & Souto, 2008; Miller et al., 2009; Miller et al., 2010). Another intriguing aspect is to examine the adaptability of vocalizations in response to different social events – ***do vocalizations evolve as new social events, such as births, occur within the group?*** (Gultekin and Hage, 2017, Gultekin and Hage, 2018; Gultekin et al., 2021) However, the studies mentioned above were limited by a small number of vocalizations and behavior examples. The automatic estimation of multiple animals' poses (Mathis et al., 2018) has been shown to be a successful approach for marmosets (Lauer et al., 2022). I hypothesize that merging this approach with our extensive marmoset detection and labeling pipeline could effectively address these questions.

3.5. Neural encoding and decoding with deep learning for vocalization processing

In the previous section (*Evolutionary origins of voice perception*), I explored future research paths to understand better how vocal perception has evolved. I suggested

extending the methods from Chapter 3, which focused on how the human brain processes voice, to study marmosets and macaques. This leads to the question: ***Would DNNs provide reasonable approximations of non-human primates' cerebral representations?*** In Chapter 2, we introduced a comprehensive dataset of marmoset vocalizations, which will be shared publicly. To effectively map these vocalizations onto cerebral representations in non-human primates, we would need sufficient brain recordings to perform brain encoding and decoding studies. However, the current datasets of brain recordings from marmosets (Jafari et al., 2023) and macaques (Bodin et al., 2021) in response to vocalizations likely contain too few stimuli. Further research is needed to overcome this limitation in the field.

4. Conclusion

To conclude, this thesis has employed artificial intelligence (AI) to investigate voice processing across primates. The synthesized model of voice processing highlights the current state-of-the-art knowledge of the workings of voice areas —and draws the limits of our knowledge.

The use of AI as a computational model has shown potential for predicting brain activity in response to voice, emphasizing the valuable role of AI in auditory neuroscience. Additionally, employing AI as a tool has aided in creating a large dataset of marmoset vocalizations, which may support future research in vocal communication and neuroethology.

Combining these multidisciplinary approaches has not only contributed to our understanding of primate auditory vocal processing but also provided some insight into the evolutionary beginnings of vocal communication. This effort has laid the groundwork for further research in auditory neuroscience, highlighting the supportive role of AI in modern neuroscience and the study of primate vocal communication.

Appendix

1. Towards studying the evolution of vocal communication systems with deep learning

Marmoset ID	Mother ID	Father ID	Date of birth	Date of death	Date of entrance	Date of exit	Sex
1	-	-	2013-12-13	-	2019-06-22	-	F
2	-	-	2013-09-23	2020-06-26	2019-06-22	2020-06-26	M
3	1	2	2018-10-22	-	2019-06-22	-	M
4	1	2	2019-03-29	-	2019-06-22	2021-08-20	F
5	-	-	2016-04-04	2022-03-07	2019-06-22	2022-03-07	M
6	-	-	2012-08-27	2020-06-18	2019-06-22	2020-06-18	F
7	6	5	2018-11-01	-	2019-06-22	-	F
8	6	5	2019-03-30	-	2019-06-22	2021-05-03	F
9	6	5	2019-03-30	-	2019-06-22	2021-05-03	F
10	-	-	2011-10-11	-	2019-06-22	-	F
11	-	-	2013-04-09	2020-01-29	2019-06-22	2020-01-29	M
12	10	11	2018-08-08	-	2019-06-22	-	F
13	10	11	2019-01-11	-	2019-06-22	-	F
14	10	11	2019-01-11	-	2019-06-22	-	F
15	1	2	2019-09-06	-	2019-09-06	2021-08-11	M
16	6	5	2019-09-06	-	2019-09-06	2021-05-03	F
17	10	11	2019-11-21	-	2019-11-21	-	M
18	10	11	2019-11-21	-	2019-11-21	-	F
19	1	2	2020-02-23	-	2020-02-23	2021-08-11	M
20	1	2	2020-02-23	-	2020-02-23	2021-09-03	M
21	1	2	2020-08-03	-	2020-08-03	2021-09-03	M

22	7	3	2020-12-07	-	2020-12-07	2023-02-03	M
23	7	3	2020-12-07	-	2020-12-07	2023-03-17	M
24	1	5	2021-02-13	-	2021-02-13	2023-01-06	F
25	1	5	2021-02-13	-	2021-02-13	-	F
26	1	5	2021-07-21	-	2021-07-21	-	F
27	1	5	2021-07-21	-	2021-07-21	-	M
28	7	3	2021-10-04	-	2021-10-04	-	M
29	7	3	2021-10-04	-	2021-10-04	2023-02-03	F
30	1	5	2022-01-26	-	2022-01-26	-	M
31	1	5	2022-01-26	-	2022-01-26	-	F
32	7	3	2022-05-02	-	2022-05-02	-	M
33	7	3	2022-05-02	-	2022-05-02	-	M
34	7	3	2022-11-04	-	2022-11-04	-	F
35	7	3	2022-11-04	-	2022-11-04	-	F

Supplementary Table S1: Description of recorded subjects. The date of entrance and exit corresponds to the period when the subject was inside the room and then recorded. Sex codes correspond to F for females and M for males.

Hyperparameter	**Value**
Sampling rate (Hz)	96,000
FFT window size	1,024
Number of frames between STFT columns (ms)	1
Reference level (dB)	20
Coefficient for pre emphasis filter	0.97
Spectral range	[125; 48,000]
Std above median to threshold out noise	1

Size in time of neighborhood-continuity filter (ms)	50
Longest distance at which two elements should be considered one (ms)	100
Smallest expected element size (in ms and Hz)	[300; 125]
Size of FFT window (ms)	4
Default dB minimum of spectrogram	-70
Threshold for spectrogram to consider noise as silence (s)	0.01
Shortest expected length of silence (s)	0.01
Longest expected vocalization (s)	5.1
Shortest expected length of syllable (s)	0.01
Threshold number of neighborhood time-frequency bins above 0 to consider a bin not noise	0.25
Size of neighborhood-continuity filter (Hz)	2,000
Proportion of temporal overlap to consider two elements one	0.25

Supplementary Table S2: Hyperparameters of the dynamic-thresholding segmentation algorithm. This algorithm is detailed in Sainburg et al., 2020 and can be accessed at https://github.com/timsainb/vocalization-segmentation.

Type	**Pre-onset (s)**	**Post-offet (s)**
Infant cry	0.1	0.3
Phee	0.15	0.5
Seep	-	0.1
Trill	-	0.4

Tsik	-	0.2
Twitter	0.5	0.5

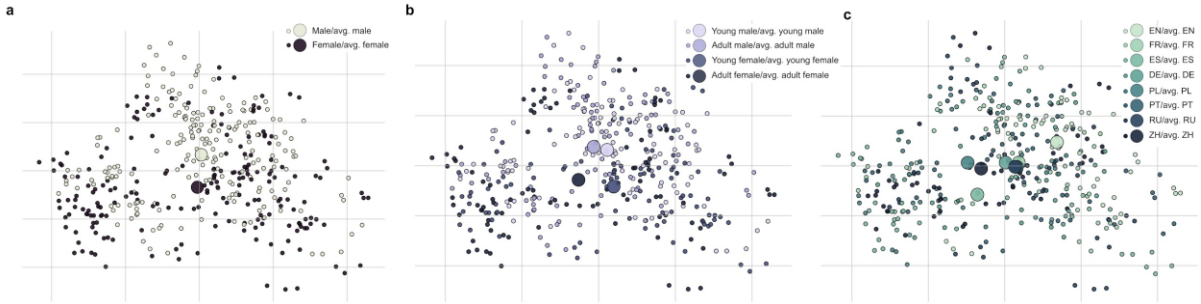
Supplementary Table S3: Description of post-classification time adjustments.

Date Type	Infant cry	Phee	Seep	Trill	Tsik	Twitter
2019-12	10.03	11.14	5.85	0.84	13.65	58.50
2020-01	36.53	19.03	1.91	3.98	2.94	35.62
2020-02	2.91	11.61	3.52	5.57	14.36	62.02
2020-03	3.72	14.13	6.40	7.52	38.41	29.82
2020-05	2.16	15.11	12.95	22.30	9.35	38.13
2020-06	6.45	12.90	6.45	40.32	0.00	33.87
2020-07	16.44	25.63	2.44	19.19	25.56	10.74
2020-09	7.01	14.92	2.18	29.34	30.51	16.04
2020-10	18.73	11.31	1.52	19.34	30.10	19.00
2020-11	9.41	10.45	2.09	20.02	36.30	21.74
2020-12	3.36	8.30	1.49	20.25	50.51	16.10
2021-01	17.52	10.09	1.14	11.31	28.45	31.49
2021-09	20.97	30.24	1.52	4.71	7.29	35.26
2022-01	1.80	21.79	2.73	9.49	17.55	46.65
2022-02	9.26	13.24	2.62	15.17	30.88	28.84
2022-03	32.28	13.60	2.14	12.12	14.21	25.65

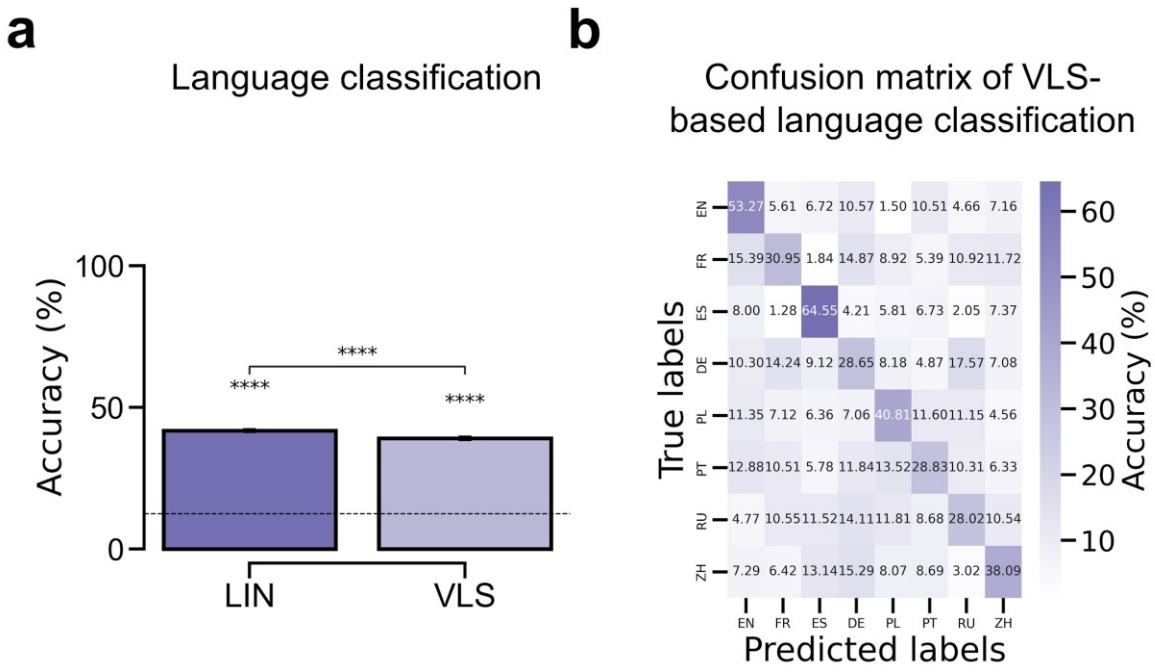
2022-04	7.33	15.58	1.68	27.13	24.59	23.69
2022-09	0.16	23.73	1.25	24.02	17.79	33.06
2022-10	0.25	24.24	1.54	27.59	16.86	29.53
2022-12	26.72	12.73	2.39	13.88	22.45	21.83
2023-02	28.40	24.12	1.08	5.54	9.09	31.78
2023-03	6.74	27.29	2.37	8.45	16.62	38.54
2023-04	4.52	29.85	2.29	9.75	14.65	38.93
Average	13.78	17.66	2.17	14.35	21.69	30.36

Supplementary Table S4: Temporal distribution of vocalizations by label over time. Distribution of 215,000 labeled vocalizations (72 hours in total). For each month, the proportion of vocalization type is indicated in %. The proportion of labeled/unlabeled vocalization is 25/75% (unlabeled omitted here).

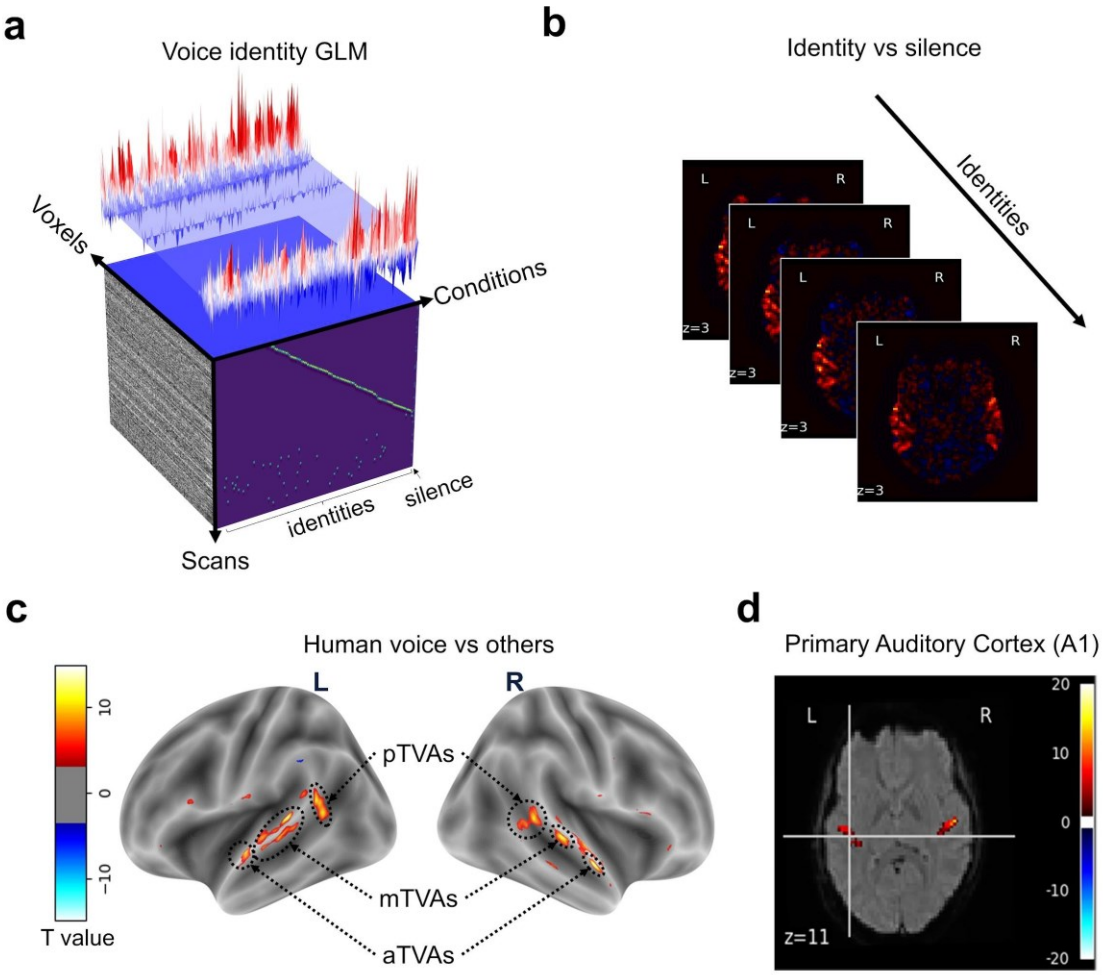
2. Encoding and decoding of voice identity in human auditory cortex



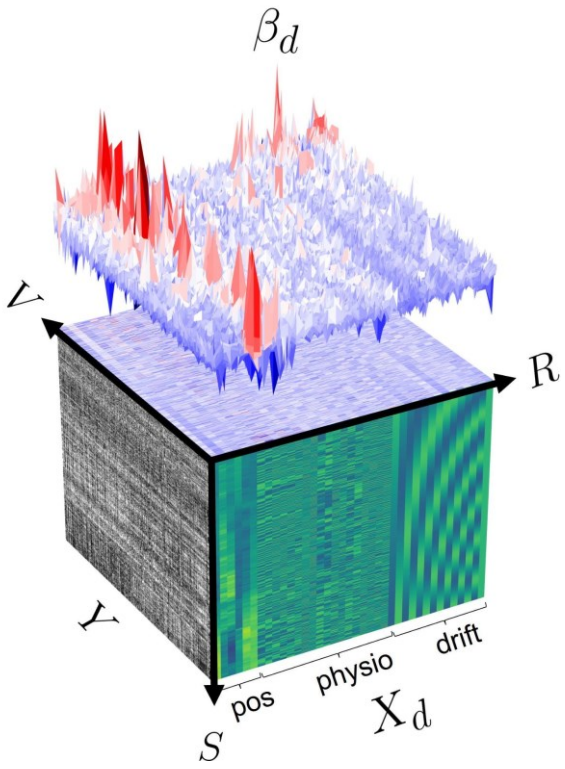
Supplementary Figure S1: Projections of the DNN-derived Voice Latent Space (VLS). Distribution of the 405 speaker identities along the first 2 principal components of the VLS coordinates from all sounds, averaged by speaker identity. Each disk represents a speaker's identity colored by either gender (as in Figure 3.1b), age, or language. **a,** Large disks represent the average of all male (black) or female (gray) speaker coordinates. ANOVAs on the first two components: PC1: $F(1, 405)=0.10$, $p=.74$; PC2: $F(1, 405)=11.00$, $p<.001$. **b,** Same for speaker age. ANOVAs on first two components: PC1: $F(1, 405)=4.12$, $p<.01$; PC2: $F(1, 405)=3.99$, $p<.01$. **c,** Same for speaker language. ANOVAs on the first two components: PC1: $F(1, 405)=8.46$, $p<.0001$; PC2: $F(1, 405)=6.09$, $p<.0001$.



Supplementary Figure S2: Language classification based on model representation. a, Performance of linear classifiers in categorizing speaker language (chance level: 12.5%) using VLS or LIN coordinates. Error bars represent the standard error of the mean (s.e.m.) across 100 random classifier initializations. All p-values are less than 1e-10. The horizontal black dashed line indicates the chance level. ****: $p < 0.0001$. **b,** Confusion matrix representing the predictions from classifiers trained on VLS features, averaged over 100 iterations.



Supplementary Figure S3: Brain activity in response to voice measured by fMRI. a, A GLM is used to model fMRI activity in response to each speaker's identity. **b,** The fMRI activity in response to each speaker's identity is mapped into dedicated voxel maps by contrasting the speaker's identity with the silence, resulting in ~135 voxel maps. **c,** The voice-sensitive ROIs used for subsequent analyses, identified in each participant via an independent Voice Localizer: the anterior, middle, and posterior Temporal Voice Areas (TVAs). **d,** The Primary Auditory Cortex (A1) is defined as the intersection between a probabilistic map of Heschl's gyri and the sound vs silence contrast map.



Supplementary Figure S4: Denoising of the fMRI BOLD responses. A general linear model (GLM) was fit to regress out the noise by predicting Y from a "denoising" design matrix X_d , composed of $R = 38$ regressors of nuisance 6 head motion parameters (3 variable for the translations, 3 variables for the rotations); 18 'RETROICOR' regressors (Glover et al., 2000) using the *TAPAS PhysIO* package (Kasper et al., 2017) with the hyperparameters set as specified in (Snoek et al., 2021); 13 regressors modeling slow artifactual trends (sines and cosines, cut-off frequency of the high-pass filter = 0.01 Hz); an intercept. The design matrix was convolved with an hemodynamic response function (HRF) with a peak at 6s sec and an undershoot at 16s sec (Glover et al., 1999), we note the convolved design matrix as $X_d \in R^{S \times R}$ where S = number of scans. The "denoise" GLM's parameters $\beta_d \in R^{R \times V}$ were optimized to minimize the amplitude of the residual $\beta_d = argmin_{\beta \in R^{R \times V}} \; || \; Y \; - X_d \, \beta \; ||^2$, where V = number of voxels. The *denoised* BOLD signal Y_d was then obtained from the original one according to $Y_d = Y - (X_d \, \beta_d)$.

Name	Layer	#Filters	Filter size	Stride	Activation
Encoder	Conv2D + BN2D	64	6x3	2x2	ReLU
	Conv2D + BN2D	128	6x2	2x2	ReLU
	Conv2D + BN2D	256	6x2	2x1	ReLU
	Conv2D + BN2D	512	6x2	2x1	ReLU
	Conv2D	7	6x2	1x1	-
Bottleneck	FC	256	-	-	-
Decoder	ConvTrans2D + BN2D	512	27x3	1x1	ReLU
	ConvTrans2D + BN2D	256	4x2	2x1	ReLU
	ConvTrans2D + BN2D	128	4x2	2x1	ReLU
	ConvTrans2D + BN2D	64	4x2	2x2	ReLU
	ConvTrans2D	1	4x2	2x2	-
Batch size	64				
Loss function	MSE + KL divergence				
Optimizer	Adam, learning rate = 0.00005				
	betas = (0.5, 0.999)				

Supplementary Table S5: Architecture of the VAE network. The architecture of the VAE consists of 15 layers with an intermediate hidden representation of 128 neurons that will stand for the VLS. The Encoder network (Enc; 7 layers) learns to map an input, s (a spectrogram of a sound), onto the (128-dimensional) VLS, while the Decoder (Dec; 7 layers) aims at reconstructing the spectrogram s from z. The learning objective of the full model is to make the output spectrogram $Dec(Enc(s))$ as close as possible to the original one s. BN: batch normalization; FC: fully connected; ReLU: Rectified Linear Unit.

Subject	ROI	Correlation	s.e.m.	T	dof	p-val	unc.	sig.	CI95%	cohen-d	BF10	power
s1	LA1	0.13 ± 0.15	0.03	4.78E+00	32	1.91E-05	****		[0.08, inf]	8.30E-01	1.22E+03	1.00
	RA1	0.21 ± 0.14	0.03	8.08E+00	32	1.57E-09	****		[0.16, inf]	1.41E+00	7.74E+06	1.00
	LmTVA	0.32 ± 0.13	0.02	1.34E+01	32	5.25E-15	****		[0.28, inf]	2.34E+00	1.27E+12	1.00
	RmTVA	0.16 ± 0.07	0.01	1.11E+01	26	1.21E-11	****		[0.13, inf]	2.13E+00	7.53E+08	1.00
	LpTVA	0.07 ± 0.13	0.02	3.15E+00	32	1.76E-03	**		[0.03, inf]	5.50E-01	2.14E+01	0.92
	RpTVA	0.04 ± 0.08	0.02	2.56E+00	31	7.82E-03	**		[0.01, inf]	4.50E-01	6.05E+00	0.80
	LaTVA	0.27 ± 0.15	0.03	1.00E+01	30	2.30E-11	****		[0.23, inf]	1.80E+00	4.20E+08	1.00
	RaTVA	0.11 ± 0.10	0.02	5.26E+00	25	9.42E-06	****		[0.07, inf]	1.03E+00	2.42E+03	1.00
	A1	0.17 ± 0.15	0.02	8.80E+00	65	5.58E-13	****		[0.14, inf]	1.08E+00	1.48E+10	1.00
	mTVA	0.25 ± 0.14	0.02	1.38E+01	59	1.71E-20	****		[0.22, inf]	1.79E+00	2.85E+17	1.00
	pTVA	0.06 ± 0.11	0.01	4.02E+00	64	7.84E-05	****		[0.03, inf]	5.00E-01	2.81E+02	0.99
	aTVA	0.20 ± 0.15	0.02	9.63E+00	56	8.92E-14	****		[0.16, inf]	1.28E+00	8.76E+10	1.00
	TVAs	0.16 ± 0.16	0.01	1.39E+01	181	8.43E-31	****		[0.14, inf]	1.03E+00	3.76E+27	1.00
s2	LA1	0.04 ± 0.11	0.02	2.16E+00	32	1.94E-02	*		[0.01, inf]	3.80E-01	2.83E+00	0.68
	RA1	-0.01 ± 0.11	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LmTVA	-0.02 ± 0.09	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RmTVA	0.03 ± 0.11	0.02	1.17E+00	21	1.27E-01	ns		[-0.01, inf]	2.50E-01	8.20E-01	0.31
	LpTVA	-0.01 ± 0.10	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RpTVA	0.04 ± 0.10	0.03	1.38E+00	16	9.37E-02	ns		[-0.01, inf]	3.30E-01	1.11E+00	0.37
	LaTVA	-0.05 ± 0.12	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RaTVA	0.03 ± 0.12	0.03	1.18E+00	19	1.26E-01	ns		[-0.01, inf]	2.60E-01	8.56E-01	0.31
	A1	0.02 ± 0.11	0.01	1.19E+00	65	1.19E-01	ns		[-0.01, inf]	1.50E-01	5.31E-01	0.32
	mTVA	0.00 ± 0.10	0.02	5.00E-02	46	4.81E-01	ns		[-0.02, inf]	1.00E-02	3.17E-01	0.06
	pTVA	0.01 ± 0.10	0.02	5.10E-01	45	3.07E-01	ns		[-0.02, inf]	7.00E-02	3.61E-01	0.13
	aTVA	-0.02 ± 0.12	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	TVAs	-0.00 ± 0.11	0.01	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
s3	LA1	0.04 ± 0.08	0.01	2.89E+00	32	3.39E-03	**		[0.02, inf]	5.00E-01	1.21E+01	0.88
	RA1	0.03 ± 0.13	0.02	1.48E+00	32	7.39E-02	ns		[-0.00, inf]	2.60E-01	1.01E+00	0.42
	LmTVA	0.04 ± 0.09	0.02	2.43E+00	28	1.10E-02	*		[0.01, inf]	4.50E-01	4.72E+00	0.76
	RmTVA	0.07 ± 0.09	0.02	4.38E+00	28	7.48E-05	****		[0.04, inf]	8.10E-01	3.64E+02	1.00
	LpTVA	0.03 ± 0.12	0.02	1.48E+00	28	7.45E-02	ns		[-0.00, inf]	2.80E-01	1.06E+00	0.42
	RpTVA	0.04 ± 0.08	0.01	2.83E+00	35	3.87E-03	**		[0.02, inf]	4.70E-01	1.05E+01	0.87
	LaTVA	0.09 ± 0.13	0.03	3.15E+00	23	2.24E-03	**		[0.04, inf]	6.40E-01	1.91E+01	0.92
	RaTVA	0.07 ± 0.12	0.03	2.41E+00	17	1.38E-02	*		[0.02, inf]	5.70E-01	4.61E+00	0.75
	A1	0.04 ± 0.11	0.01	2.80E+00	65	3.38E-03	**		[0.01, inf]	3.40E-01	9.50E+00	0.87
	mTVA	0.06 ± 0.09	0.01	4.76E+00	57	6.83E-06	****		[0.04, inf]	6.30E-01	2.76E+03	1.00
	pTVA	0.04 ± 0.10	0.01	2.92E+00	64	2.40E-03	**		[0.02, inf]	3.60E-01	1.29E+01	0.89
	aTVA	0.08 ± 0.13	0.02	4.00E+00	41	1.28E-04	***		[0.05, inf]	6.20E-01	2.05E+02	0.99
	TVAs	0.06 ± 0.11	0.01	6.62E+00	164	2.46E-10	****		[0.04, inf]	5.20E-01	3.49E+07	1.00
all	LA1	0.07 ± 0.12	0.01	5.58E+00	98	1.05E-07	****		[0.05, inf]	5.60E-01	1.21E+05	1.00
	RA1	0.08 ± 0.16	0.02	4.82E+00	98	2.60E-06	****		[0.05, inf]	4.80E-01	5.85E+03	1.00
	LmTVA	0.13 ± 0.19	0.02	6.37E+00	86	4.45E-09	****		[0.10, inf]	6.80E-01	2.55E+06	1.00
	RmTVA	0.09 ± 0.10	0.01	7.55E+00	77	3.72E-11	****		[0.07, inf]	8.50E-01	2.55E+08	1.00
	LpTVA	0.03 ± 0.12	0.01	2.66E+00	90	4.59E-03	**		[0.01, inf]	2.80E-01	6.39E+00	0.84
	RpTVA	0.04 ± 0.09	0.01	4.01E+00	84	6.63E-05	****		[0.02, inf]	4.30E-01	3.00E+02	0.99
	LaTVA	0.11 ± 0.19	0.02	5.07E+00	83	1.20E-06	****		[0.07, inf]	5.50E-01	1.27E+04	1.00
	RaTVA	0.07 ± 0.12	0.01	5.01E+00	63	2.34E-06	****		[0.05, inf]	6.30E-01	7.30E+03	1.00
	A1	0.07 ± 0.14	0.01	7.25E+00	197	4.67E-12	****		[0.06, inf]	5.20E-01	1.54E+09	1.00
	mTVA	0.11 ± 0.16	0.01	9.12E+00	164	1.30E-16	****		[0.09, inf]	7.10E-01	4.40E+13	1.00
	pTVA	0.04 ± 0.11	0.01	4.49E+00	175	6.53E-06	****		[0.02, inf]	3.40E-01	2.00E+03	1.00
	aTVA	0.09 ± 0.17	0.01	6.81E+00	147	1.14E-10	****		[0.07, inf]	5.60E-01	7.58E+07	1.00
	TVAs	0.08 ± 0.15	0.01	1.18E+01	488	9.58E-29	****		[0.07, inf]	5.30E-01	2.93E+25	1.00

Supplementary Table S6: Assessing the significance of brain encoding performance with LIN features. This table reports the significance of the brain encoding performance with LIN features. We compared the distribution of Pearson's correlation coefficients to the chance level of 0.0 by conducting one-sample t-tests. Using a linear model, we calculated the correlation between the voxels in the speaker activity maps and the predicted voxels from the LIN features. s.e.m. = standard error of the mean. all = we combined the scores of all participants before computing the test. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each participant and ROI.

Subject	ROI	Correlation	s.e.m.	T	dof	p-val	unc.	sig.	CI95%	cohen-d	BF10	power
s1	LA1	0.03 ± 0.11	0.02	1.46E+00	32	7.71E-02	ns		[-0.00, inf]	2.50E-01	9.75E-01	0.41
	RA1	0.13 ± 0.09	0.02	8.06E+00	32	1.67E-09	****		[0.10, inf]	1.40E+00	7.28E+06	1.00
	LmTVA	0.25 ± 0.16	0.03	8.95E+00	32	1.58E-10	****		[0.20, inf]	1.56E+00	6.77E+07	1.00
	RmTVA	0.08 ± 0.09	0.02	4.89E+00	26	2.24E-05	****		[0.05, inf]	9.40E-01	1.09E+03	1.00
	LpTVA	-0.03 ± 0.12	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RpTVA	-0.06 ± 0.11	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LaTVA	0.15 ± 0.16	0.03	5.34E+00	30	4.43E-06	****		[0.10, inf]	9.60E-01	4.70E+03	1.00
	RaTVA	0.03 ± 0.11	0.02	1.55E+00	25	6.70E-02	ns		[-0.00, inf]	3.00E-01	1.19E+00	0.44
	A1	0.08 ± 0.11	0.01	5.65E+00	65	1.93E-07	****		[0.06, inf]	7.00E-01	7.57E+04	1.00
	mTVA	0.17 ± 0.15	0.02	8.69E+00	59	1.91E-12	****		[0.14, inf]	1.12E+00	4.57E+09	1.00
	pTVA	-0.04 ± 0.12	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	aTVA	0.10 ± 0.15	0.02	4.94E+00	56	3.68E-06	****		[0.07, inf]	6.50E-01	4.93E+03	1.00
	TVAs	0.07 ± 0.17	0.01	5.79E+00	181	1.52E-08	****		[0.05, inf]	4.30E-01	6.37E+05	1.00
s2	LA1	0.04 ± 0.14	0.02	1.51E+00	32	7.01E-02	ns		[-0.00, inf]	2.60E-01	1.05E+00	0.43
	RA1	0.01 ± 0.12	0.02	3.60E-01	32	3.59E-01	ns		[-0.03, inf]	6.00E-02	3.96E-01	0.10
	LmTVA	0.04 ± 0.07	0.01	3.07E+00	24	2.61E-03	**		[0.02, inf]	6.10E-01	1.66E+01	0.91
	RmTVA	0.08 ± 0.10	0.02	3.95E+00	21	3.64E-04	***		[0.05, inf]	8.40E-01	9.46E+01	0.99
	LpTVA	-0.01 ± 0.10	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RpTVA	0.02 ± 0.13	0.03	7.30E-01	16	2.39E-01	ns		[-0.03, inf]	1.80E-01	6.29E-01	0.17
	LaTVA	-0.01 ± 0.08	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RaTVA	0.02 ± 0.08	0.02	1.00E+00	19	1.64E-01	ns		[-0.01, inf]	2.20E-01	7.26E-01	0.25
	A1	0.02 ± 0.13	0.02	1.38E+00	65	8.61E-02	ns		[-0.00, inf]	1.70E-01	6.66E-01	0.39
	mTVA	0.06 ± 0.09	0.01	4.92E+00	46	5.72E-06	****		[0.04, inf]	7.20E-01	3.43E+03	1.00
	pTVA	-0.00 ± 0.11	0.02	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	aTVA	0.00 ± 0.08	0.01	4.10E-01	48	3.43E-01	ns		[-0.02, inf]	6.00E-02	3.36E-01	0.11
	TVAs	0.02 ± 0.10	0.01	2.65E+00	141	4.46E-03	**		[0.01, inf]	2.20E-01	5.41E+00	0.84
s3	LA1	0.01 ± 0.09	0.02	3.50E-01	32	3.66E-01	ns		[-0.02, inf]	6.00E-02	3.94E-01	0.10
	RA1	0.03 ± 0.11	0.02	1.62E+00	32	5.78E-02	ns		[-0.00, inf]	2.80E-01	1.21E+00	0.48
	LmTVA	0.05 ± 0.14	0.03	2.03E+00	28	2.61E-02	*		[0.01, inf]	3.80E-01	2.34E+00	0.63
	RmTVA	0.09 ± 0.08	0.02	5.64E+00	28	2.41E-06	****		[0.06, inf]	1.05E+00	8.29E+03	1.00
	LpTVA	0.00 ± 0.10	0.02	2.20E-01	28	4.12E-01	ns		[-0.03, inf]	4.00E-02	4.04E-01	0.08
	RpTVA	0.01 ± 0.11	0.02	4.50E-01	35	3.30E-01	ns		[-0.02, inf]	7.00E-02	3.93E-01	0.11
	LaTVA	0.04 ± 0.12	0.03	1.60E+00	23	6.16E-02	ns		[-0.00, inf]	3.30E-01	1.31E+00	0.46
	RaTVA	0.11 ± 0.12	0.03	3.65E+00	17	9.96E-04	***		[0.06, inf]	8.60E-01	4.13E+01	0.97
	A1	0.02 ± 0.10	0.01	1.49E+00	65	7.09E-02	ns		[-0.00, inf]	1.80E-01	7.69E-01	0.43
	mTVA	0.07 ± 0.11	0.02	4.68E+00	57	9.11E-06	****		[0.05, inf]	6.10E-01	2.12E+03	1.00
	pTVA	0.01 ± 0.11	0.01	4.90E-01	64	3.14E-01	ns		[-0.02, inf]	6.00E-02	3.05E-01	0.12
	aTVA	0.07 ± 0.13	0.02	3.53E+00	41	5.14E-04	***		[0.04, inf]	5.50E-01	5.87E+01	0.97
	TVAs	0.05 ± 0.12	0.01	4.87E+00	164	1.32E-06	****		[0.03, inf]	3.80E-01	9.31E+03	1.00
all	LA1	0.02 ± 0.11	0.01	2.04E+00	98	2.19E-02	*		[0.00, inf]	2.10E-01	1.62E+00	0.65
	RA1	0.06 ± 0.12	0.01	4.67E+00	98	4.87E-06	****		[0.04, inf]	4.70E-01	3.24E+03	1.00
	LmTVA	0.12 ± 0.16	0.02	7.09E+00	86	1.77E-10	****		[0.09, inf]	7.60E-01	5.59E+07	1.00
	RmTVA	0.09 ± 0.09	0.01	8.47E+00	77	6.43E-13	****		[0.07, inf]	9.60E-01	1.27E+10	1.00
	LpTVA	-0.01 ± 0.11	0.01	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RpTVA	-0.02 ± 0.12	0.01	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LaTVA	0.07 ± 0.14	0.02	4.23E+00	83	2.96E-05	****		[0.04, inf]	4.60E-01	6.36E+02	0.99
	RaTVA	0.05 ± 0.11	0.01	3.57E+00	63	3.50E-04	***		[0.03, inf]	4.50E-01	7.23E+01	0.97
	A1	0.04 ± 0.12	0.01	4.76E+00	197	1.89E-06	****		[0.03, inf]	3.40E-01	6.19E+03	1.00
	mTVA	0.11 ± 0.13	0.01	1.01E+01	164	2.66E-19	****		[0.09, inf]	7.90E-01	1.88E+16	1.00
	pTVA	-0.01 ± 0.12	0.01	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	aTVA	0.06 ± 0.13	0.01	5.52E+00	147	7.61E-08	****		[0.04, inf]	4.50E-01	1.46E+05	1.00
	TVAs	0.05 ± 0.14	0.01	7.88E+00	488	1.05E-14	****		[0.04, inf]	3.60E-01	4.33E+11	1.00

Supplementary Table S7: Assessing the significance of brain encoding performance with VLS features. This table reports the significance of the brain encoding performance with VLS features. We compared the distribution of Pearson's correlation coefficients to the chance level of 0.0 by conducting one-sample t-tests. Using a linear model, we calculated the correlation between the voxels in the speaker activity maps and the predicted voxels from the VLS features. s.e.m. = standard error of the mean. all = we combined the scores of all participants before computing the test. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each participant and ROI.

Subject	ROI	Correlation VLS	Correlation LIN	s.e.m. VLS	s.e.m. LIN	T VLS vs LIN	dof	p-val	unc. sig.	CI95%	cohen-d	BF10	power
s1	LA1	0.03 ± 0.11	0.13 ± 0.15	0.02	0.03	-4.43E+00	32	1.03E-04	***	[-0.14, -0.05]	7.30E-01	2.47E+02	0.98
	RA1	0.13 ± 0.09	0.21 ± 0.14	0.02	0.03	-3.75E+00	32	7.07E-04	***	[-0.11, -0.03]	6.00E-01	4.39E+01	0.92
	LmTVA	0.25 ± 0.16	0.32 ± 0.13	0.03	0.02	-3.90E+00	32	4.61E-04	***	[-0.11, -0.03]	4.80E-01	6.43E+01	0.76
	RmTVA	0.08 ± 0.09	0.16 ± 0.07	0.02	0.01	-5.48E+00	26	9.54E-06	****	[-0.10, -0.05]	9.20E-01	2.24E+03	1.00
	LpTVA	-0.03 ± 0.12	0.07 ± 0.13	0.02	0.02	-6.49E+00	32	2.68E-07	****	[-0.13, -0.07]	7.60E-01	5.95E+04	0.99
	RpTVA	-0.06 ± 0.11	0.04 ± 0.08	0.02	0.02	-5.09E+00	31	1.67E-05	****	[-0.14, -0.06]	1.01E+00	1.31E+03	1.00
	LaTVA	0.15 ± 0.16	0.27 ± 0.15	0.03	0.03	-7.34E+00	30	3.55E-08	****	[-0.15, -0.09]	7.70E-01	3.95E+05	0.99
	RaTVA	0.03 ± 0.11	0.11 ± 0.10	0.02	0.02	-4.24E+00	25	2.65E-04	***	[-0.11, -0.04]	7.10E-01	1.11E+02	0.93
	A1	0.08 ± 0.11	0.17 ± 0.15	0.01	0.02	-5.81E+00	65	2.02E-07	****	[-0.12, -0.06]	6.30E-01	6.96E+04	1.00
	mTVA	0.17 ± 0.15	0.25 ± 0.14	0.02	0.02	-6.24E+00	59	5.16E-08	****	[-0.10, -0.05]	5.00E-01	2.58E+05	0.97
	pTVA	-0.04 ± 0.12	0.06 ± 0.11	0.02	0.01	-8.06E+00	64	2.58E-11	****	[-0.12, -0.08]	8.60E-01	3.62E+08	1.00
	aTVA	0.10 ± 0.15	0.20 ± 0.15	0.02	0.02	-8.11E+00	56	5.09E-11	****	[-0.13, -0.08]	6.60E-01	1.91E+08	1.00
	TVAs	0.07 ± 0.17	0.16 ± 0.16	0.01	0.01	-1.29E+01	181	1.85E-27	****	[-0.11, -0.08]	5.60E-01	1.89E+24	1.00
s2	LA1	0.04 ± 0.14	0.04 ± 0.11	0.02	0.02	-2.70E-01	32	7.93E-01	ns	[-0.03, 0.02]	3.00E-02	1.92E-01	0.05
	RA1	0.01 ± 0.12	-0.01 ± 0.11	0.02	0.02	6.20E-01	32	5.38E-01	ns	[-0.03, 0.06]	1.30E-01	2.23E-01	0.11
	LmTVA	0.04 ± 0.07	-0.02 ± 0.09	0.01	0.02	3.52E+00	24	1.77E-03	**	[0.03, 0.11]	8.10E-01	2.11E+01	0.97
	RmTVA	0.08 ± 0.10	0.03 ± 0.11	0.02	0.02	3.74E+00	21	1.22E-03	**	[0.02, 0.09]	5.20E-01	3.01E+01	0.65
	LpTVA	-0.01 ± 0.10	-0.01 ± 0.10	0.02	0.02	-4.20E-01	28	6.78E-01	ns	[-0.04, 0.02]	6.00E-02	2.14E-01	0.06
	RpTVA	0.02 ± 0.13	0.04 ± 0.10	0.03	0.03	-4.10E-01	16	6.88E-01	ns	[-0.07, 0.05]	1.00E-01	2.68E-01	0.07
	LaTVA	-0.01 ± 0.08	-0.05 ± 0.12	0.02	0.02	2.78E+00	28	9.51E-03	**	[0.01, 0.08]	4.70E-01	4.75E+00	0.68
	RaTVA	0.02 ± 0.08	0.03 ± 0.12	0.02	0.03	-4.30E-01	19	6.69E-01	ns	[-0.07, 0.05]	1.20E-01	2.53E-01	0.08
	A1	0.02 ± 0.13	0.02 ± 0.11	0.02	0.01	4.00E-01	65	6.87E-01	ns	[-0.02, 0.03]	5.00E-02	1.46E-01	0.07
	mTVA	0.06 ± 0.09	0.00 ± 0.10	0.01	0.02	5.06E+00	46	7.24E-06	****	[0.04, 0.09]	6.40E-01	2.62E+03	0.99
	pTVA	-0.00 ± 0.11	0.01 ± 0.10	0.02	0.02	-5.90E-01	45	5.57E-01	ns	[-0.04, 0.02]	8.00E-02	1.89E-01	0.08
	aTVA	0.00 ± 0.08	-0.02 ± 0.12	0.01	0.02	1.50E+00	48	1.40E-01	ns	[-0.01, 0.06]	2.20E-01	4.43E-01	0.33
	TVAs	0.02 ± 0.10	-0.00 ± 0.11	0.01	0.01	3.06E+00	141	2.64E-03	**	[0.01, 0.04]	2.40E-01	8.00E+00	0.83
s3	LA1	0.01 ± 0.09	0.04 ± 0.08	0.02	0.01	-2.32E+00	32	2.68E-02	*	[-0.07, -0.00]	4.20E-01	1.91E+00	0.64
	RA1	0.03 ± 0.11	0.03 ± 0.13	0.02	0.02	-1.00E-01	32	9.17E-01	ns	[-0.04, 0.03]	1.00E-02	1.87E-01	0.05
	LmTVA	0.05 ± 0.14	0.04 ± 0.09	0.03	0.02	7.20E-01	28	4.79E-01	ns	[-0.02, 0.04]	8.00E-02	2.50E-01	0.07
	RmTVA	0.09 ± 0.08	0.07 ± 0.09	0.02	0.02	9.30E-01	28	3.59E-01	ns	[-0.02, 0.05]	1.80E-01	2.94E-01	0.16
	LpTVA	0.00 ± 0.10	0.03 ± 0.12	0.02	0.02	-1.82E+00	28	7.91E-02	ns	[-0.06, 0.00]	2.50E-01	8.47E-01	0.26
	RpTVA	0.01 ± 0.11	0.04 ± 0.08	0.02	0.01	-2.26E+00	35	3.03E-02	*	[-0.06, -0.00]	3.10E-01	1.67E+00	0.44
	LaTVA	0.04 ± 0.12	0.09 ± 0.13	0.03	0.03	-3.71E+00	23	1.15E-03	**	[-0.07, -0.02]	3.70E-01	3.10E+01	0.40
	RaTVA	0.11 ± 0.12	0.07 ± 0.12	0.03	0.03	2.79E+00	17	1.25E-02	*	[0.01, 0.07]	3.00E-01	4.41E+00	0.23
	A1	0.02 ± 0.10	0.04 ± 0.11	0.01	0.01	-1.60E+00	65	1.14E-01	ns	[-0.04, 0.00]	1.80E-01	4.55E-01	0.29
	mTVA	0.07 ± 0.11	0.06 ± 0.09	0.02	0.01	1.19E+00	57	2.40E-01	ns	[-0.01, 0.03]	1.20E-01	2.79E-01	0.15
	pTVA	0.01 ± 0.11	0.04 ± 0.10	0.01	0.01	-2.92E+00	64	4.88E-03	**	[-0.05, -0.01]	2.80E-01	6.36E+00	0.61
	aTVA	0.07 ± 0.13	0.08 ± 0.13	0.02	0.02	-9.50E-01	41	3.49E-01	ns	[-0.03, 0.01]	8.00E-02	2.54E-01	0.08
	TVAs	0.05 ± 0.12	0.06 ± 0.11	0.01	0.01	-1.54E+00	164	1.25E-01	ns	[-0.02, 0.00]	9.00E-02	2.77E-01	0.20
all	LA1	0.02 ± 0.11	0.07 ± 0.12	0.01	0.01	-4.25E+00	98	4.92E-05	****	[-0.07, -0.02]	3.80E-01	3.57E+02	0.97
	RA1	0.06 ± 0.12	0.08 ± 0.16	0.01	0.02	-1.64E+00	98	1.04E-01	ns	[-0.04, 0.00]	1.40E-01	4.06E-01	0.29
	LmTVA	0.12 ± 0.16	0.13 ± 0.19	0.02	0.02	-4.10E-01	86	6.80E-01	ns	[-0.03, 0.02]	3.00E-02	1.29E-01	0.06
	RmTVA	0.09 ± 0.09	0.09 ± 0.10	0.01	0.01	-4.00E-01	77	6.87E-01	ns	[-0.03, 0.02]	4.00E-02	1.35E-01	0.07
	LpTVA	-0.01 ± 0.11	0.03 ± 0.12	0.01	0.01	-4.81E+00	90	5.96E-06	****	[-0.07, -0.03]	4.00E-01	2.64E+03	0.96
	RpTVA	-0.02 ± 0.12	0.04 ± 0.09	0.01	0.01	-4.60E+00	84	1.51E-05	****	[-0.08, -0.03]	5.10E-01	1.13E+03	1.00
	LaTVA	0.07 ± 0.14	0.11 ± 0.19	0.02	0.02	-3.42E+00	83	9.61E-04	***	[-0.07, -0.02]	2.40E-01	2.46E+01	0.59
	RaTVA	0.05 ± 0.11	0.07 ± 0.12	0.01	0.01	-1.81E+00	63	7.58E-02	ns	[-0.05, 0.00]	2.10E-01	6.31E-01	0.37
	A1	0.04 ± 0.12	0.07 ± 0.14	0.01	0.01	-4.02E+00	197	8.19E-05	****	[-0.05, -0.02]	2.50E-01	1.70E+02	0.94
	mTVA	0.11 ± 0.13	0.11 ± 0.16	0.01	0.01	-5.80E-01	164	5.64E-01	ns	[-0.02, 0.01]	3.00E-02	1.02E-01	0.07
	pTVA	-0.01 ± 0.12	0.04 ± 0.11	0.01	0.01	-6.65E+00	175	3.68E-10	****	[-0.06, -0.04]	4.50E-01	2.27E+07	1.00
	aTVA	0.06 ± 0.13	0.09 ± 0.17	0.01	0.01	-3.79E+00	147	2.23E-04	***	[-0.05, -0.02]	2.30E-01	7.55E+01	0.78
	TVAs	0.05 ± 0.14	0.08 ± 0.15	0.01	0.01	-6.28E+00	488	7.44E-10	****	[-0.04, -0.02]	2.10E-01	7.90E+06	1.00

Supplementary Table S8: Comparing the performance of brain encoding models. This table reports the significance of the VLS-LIN difference in the brain encoding performance. We conducted paired t-tests between the brain encoding model's scores trained with the VLS features to predict the speaker activity maps' voxels and those trained with the LIN features. s.e.m. = standard error of the mean. all = we combined the scores of all participants before computing the test. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each participant and ROI.

Subject	Model	ROI	Correlation ROI	Correlation A1	s.e.m. ROI	s.e.m. A1	T ROI vs A1	dof	p-val	unc. sig.	CI95%	cohen-d	BF10	power
s1	LIN	mTVA	0.25 ± 0.14	0.17 ± 0.15	0.02	0.02	3.070000	124	2.62E-03	**	[0.03, 0.13]	5.50E-01	1.25E+01	0.86
		pTVA	0.06 ± 0.11	0.17 ± 0.15	0.01	0.02	-4.710000	129	6.34E-06	****	[-0.16, -0.06]	8.20E-01	2.60E+03	1.00
		aTVA	0.20 ± 0.15	0.17 ± 0.15	0.02	0.02	1.150000	121	2.53E-01	ns	[-0.02, 0.09]	2.10E-01	3.50E-01	0.21
		TVAs	0.16 ± 0.16	0.17 ± 0.15	0.01	0.02	-0.130000	246	8.93E-01	ns	[-0.05, 0.04]	2.00E-02	1.57E-01	0.05
	VLS	mTVA	0.17 ± 0.15	0.08 ± 0.11	0.02	0.01	3.860000	124	1.81E-04	***	[0.05, 0.14]	6.90E-01	1.28E+02	0.97
		pTVA	-0.04 ± 0.12	0.08 ± 0.11	0.02	0.01	-6.020000	129	1.68E-08	****	[-0.17, -0.08]	1.05E+00	6.23E+05	1.00
		aTVA	0.10 ± 0.15	0.08 ± 0.11	0.02	0.01	0.750000	121	4.53E-01	ns	[-0.03, 0.07]	1.40E-01	2.49E-01	0.12
		TVAs	0.07 ± 0.17	0.08 ± 0.11	0.01	0.01	-0.350000	246	7.25E-01	ns	[-0.05, 0.04]	5.00E-02	1.65E-01	0.06
s2	LIN	mTVA	0.00 ± 0.10	0.02 ± 0.11	0.02	0.01	-0.760000	111	4.48E-01	ns	[-0.06, 0.03]	1.50E-01	2.62E-01	0.12
		pTVA	0.01 ± 0.10	0.02 ± 0.11	0.02	0.01	-0.430000	110	6.70E-01	ns	[-0.05, 0.03]	8.00E-02	2.21E-01	0.07
		aTVA	-0.02 ± 0.12	0.02 ± 0.11	0.02	0.01	-1.580000	113	1.16E-01	ns	[-0.08, 0.01]	3.00E-01	6.15E-01	0.35
		TVAs	-0.00 ± 0.11	0.02 ± 0.11	0.01	0.01	-1.220000	206	2.22E-01	ns	[-0.05, 0.01]	1.80E-01	3.24E-01	0.23
	VLS	mTVA	0.06 ± 0.09	0.02 ± 0.13	0.01	0.02	1.810000	111	7.29E-02	ns	[-0.00, 0.08]	3.50E-01	8.70E-01	0.43
		pTVA	-0.00 ± 0.11	0.02 ± 0.13	0.02	0.02	-0.960000	110	3.41E-01	ns	[-0.07, 0.02]	1.80E-01	3.06E-01	0.16
		aTVA	0.00 ± 0.08	0.02 ± 0.13	0.01	0.02	-0.810000	113	4.20E-01	ns	[-0.06, 0.03]	1.50E-01	2.69E-01	0.13
		TVAs	0.02 ± 0.10	0.02 ± 0.13	0.01	0.02	-0.020000	206	9.87E-01	ns	[-0.03, 0.03]	0.00E+00	1.62E-01	0.05
s3	LIN	mTVA	0.06 ± 0.09	0.04 ± 0.11	0.01	0.01	1.170000	122	2.43E-01	ns	[-0.01, 0.06]	2.10E-01	3.57E-01	0.21
		pTVA	0.04 ± 0.10	0.04 ± 0.11	0.01	0.01	-0.040000	129	9.71E-01	ns	[-0.04, 0.03]	1.00E-02	1.87E-01	0.05
		aTVA	0.08 ± 0.13	0.04 ± 0.11	0.02	0.01	1.940000	106	5.55E-02	ns	[-0.00, 0.09]	3.80E-01	1.09E+00	0.48
		TVAs	0.06 ± 0.11	0.04 ± 0.11	0.01	0.01	1.190000	229	2.35E-01	ns	[-0.01, 0.05]	1.70E-01	3.06E-01	0.22
	VLS	mTVA	0.07 ± 0.11	0.02 ± 0.10	0.02	0.01	2.700000	122	7.97E-03	**	[0.01, 0.09]	4.90E-01	4.89E+00	0.76
		pTVA	0.01 ± 0.11	0.02 ± 0.10	0.01	0.01	-0.650000	129	5.16E-01	ns	[-0.05, 0.02]	1.10E-01	2.27E-01	0.10
		aTVA	0.07 ± 0.13	0.02 ± 0.10	0.02	0.01	2.340000	106	2.11E-02	*	[0.01, 0.10]	4.60E-01	2.32E+00	0.64
		TVAs	0.05 ± 0.12	0.02 ± 0.10	0.01	0.01	1.610000	229	1.08E-01	ns	[-0.01, 0.06]	2.30E-01	5.30E-01	0.36
all	LIN	mTVA	0.11 ± 0.16	0.07 ± 0.14	0.01	0.01	2.360000	361	1.86E-02	*	[0.01, 0.07]	2.50E-01	1.69E+00	0.65
		pTVA	0.04 ± 0.11	0.07 ± 0.14	0.01	0.01	-2.850000	372	4.57E-03	**	[-0.06, -0.01]	3.00E-01	5.59E+00	0.81
		aTVA	0.09 ± 0.17	0.07 ± 0.14	0.01	0.01	1.200000	344	2.29E-01	ns	[-0.01, 0.05]	1.30E-01	2.40E-01	0.22
		TVAs	0.08 ± 0.15	0.07 ± 0.14	0.01	0.01	0.410000	685	6.79E-01	ns	[-0.02, 0.03]	3.00E-02	1.02E-01	0.07
	VLS	mTVA	0.11 ± 0.13	0.04 ± 0.12	0.01	0.01	4.910000	361	1.40E-06	****	[0.04, 0.09]	5.20E-01	9.29E+03	1.00
		pTVA	-0.01 ± 0.12	0.04 ± 0.12	0.01	0.01	-4.450000	372	1.13E-05	****	[-0.08, -0.03]	4.60E-01	1.31E+03	0.99
		aTVA	0.06 ± 0.13	0.04 ± 0.12	0.01	0.01	1.410000	344	1.58E-01	ns	[-0.01, 0.05]	1.50E-01	3.13E-01	0.29
		TVAs	0.05 ± 0.14	0.04 ± 0.12	0.01	0.01	0.750000	685	4.56E-01	ns	[-0.01, 0.03]	6.00E-02	1.23E-01	0.12

Supplementary Table S9: Comparing the performance of brain encoding ROIs. This table reports the significance of the A1-TVAs difference in the brain encoding performance. We conducted two-sample t-tests between the brain encoding model's scores trained to predict A1 and those trained to predict temporal voice areas. s.e.m. = standard error of the mean. all = we combined the scores of all participants before computing the test. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each participant and model.

Subject	Model	ROI	Correlation	p-unc	p-corr	corr. sig.
s1	LIN	LA1	0.07	1.39E-02	1.79E-01	ns
		RA1	0.08	4.20E-03	1.04E-01	ns
		LmTVA	0.08	1.92E-02	3.80E-01	ns
		RmTVA	0.06	7.14E-02	5.51E-01	ns
		LpTVA	0.04	2.05E-01	6.53E-01	ns
		RpTVA	0.03	3.16E-01	7.66E-01	ns
		LaTVA	0.12	1.40E-03	5.04E-01	ns
		RaTVA	0.07	4.26E-02	6.61E-01	ns
	VLS	LA1	0.09	6.52E-02	6.53E-02	ns
		RA1	0.08	7.47E-02	7.49E-02	ns
		LmTVA	0.11	1.28E-01	1.28E-01	ns
		RmTVA	0.10	1.39E-01	1.39E-01	ns
		LpTVA	0.09	1.94E-01	1.94E-01	ns
		RpTVA	0.11	1.18E-01	1.18E-01	ns
		LaTVA	0.19	4.17E-02	4.17E-02	*
		RaTVA	0.13	1.30E-01	1.30E-01	ns
s2	LIN	LA1	-0.01	5.03E-01	6.27E-01	ns
		RA1	-0.00	1.87E-01	4.95E-01	ns
		LmTVA	0.01	4.72E-01	7.19E-01	ns
		RmTVA	-0.01	7.98E-01	9.03E-01	ns
		LpTVA	-0.01	7.21E-01	8.13E-01	ns
		RpTVA	0.00	4.07E-01	6.00E-01	ns
		LaTVA	-0.02	8.62E-01	9.22E-01	ns
		RaTVA	-0.02	7.69E-01	7.92E-01	ns
	VLS	LA1	0.02	2.36E-01	2.52E-01	ns
		RA1	0.03	1.12E-01	1.12E-01	ns
		LmTVA	0.06	2.29E-01	2.29E-01	ns
		RmTVA	-0.01	8.59E-01	9.26E-01	ns
		LpTVA	-0.02	8.89E-01	9.85E-01	ns
		RpTVA	0.01	4.22E-01	4.54E-01	ns
		LaTVA	0.03	3.37E-01	3.38E-01	ns
		RaTVA	0.00	2.76E-01	3.23E-01	ns
s3	LIN	LA1	-0.00	5.71E-01	6.66E-01	ns
		RA1	0.05	3.00E-04	5.00E-02	*
		LmTVA	0.05	4.10E-03	2.04E-01	ns
		RmTVA	0.05	2.20E-03	1.16E-01	ns
		LpTVA	0.05	5.80E-03	1.73E-01	ns
		RpTVA	0.04	2.66E-02	4.60E-01	ns
		LaTVA	0.12	0.00E+00	7.70E-02	ns
		RaTVA	0.03	3.35E-02	3.26E-01	ns
	VLS	LA1	0.02	1.78E-01	2.10E-01	ns
		RA1	0.07	1.42E-02	1.42E-02	*
		LmTVA	0.11	7.20E-03	7.20E-03	**
		RmTVA	0.05	1.23E-01	1.23E-01	ns
		LpTVA	0.08	5.82E-02	5.82E-02	ns
		RpTVA	0.13	1.56E-02	1.56E-02	*
		LaTVA	0.23	1.00E-04	1.00E-04	****
		RaTVA	0.04	2.34E-01	2.34E-01	ns

Supplementary Table S10: Assessing the significance of the RSA brain-model correlation. This table reports the significance of the RSA brain-model performance. The brain-model correlation coefficients were computed between the ranked representational dissimilarity matrices. The correlation was compared to 0 using a 'maximum statistics' approach in which they are compared to a distribution of correlation coefficients drawn from a large number of random permutations of the model RDMs' rows and columns while controlling for the number of comparisons performed (cf. Methods) (Maris & Oostenveld, 2007), for each participant, model and ROI.

Subject	ROI	Correlation VLS	Correlation LIN	p-corr	p-unc	corr. sig.
s1	LA1	0.09	0.07	4.45E-01	2.99E-01	ns
	RA1	0.08	0.08	8.30E-01	5.05E-01	ns
	LmTVA	0.11	0.08	4.63E-01	4.51E-01	ns
	RmTVA	0.10	0.06	3.98E-01	3.97E-01	ns
	LpTVA	0.09	0.04	2.86E-01	2.84E-01	ns
	RpTVA	0.11	0.03	1.11E-01	1.11E-01	ns
	LaTVA	0.19	0.12	3.94E-01	3.94E-01	ns
	RaTVA	0.13	0.07	3.48E-01	3.48E-01	ns
s2	LA1	0.02	-0.01	3.25E-01	1.65E-01	ns
	RA1	0.03	-0.00	1.58E-01	1.41E-01	ns
	LmTVA	0.06	0.01	1.78E-01	1.72E-01	ns
	RmTVA	-0.01	-0.01	1.00E+00	8.15E-01	ns
	LpTVA	-0.02	-0.01	1.00E+00	8.72E-01	ns
	RpTVA	0.01	0.00	7.13E-01	4.47E-01	ns
	LaTVA	0.03	-0.02	1.20E-01	1.19E-01	ns
	RaTVA	0.00	-0.02	3.94E-01	1.13E-01	ns
s3	LA1	0.02	-0.00	3.22E-01	1.05E-01	ns
	RA1	0.07	0.05	4.83E-01	3.22E-01	ns
	LmTVA	0.11	0.05	6.61E-02	6.25E-02	ns
	RmTVA	0.05	0.05	1.00E+00	5.38E-01	ns
	LpTVA	0.08	0.05	4.30E-01	3.08E-01	ns
	RpTVA	0.13	0.04	3.66E-02	3.66E-02	*
	LaTVA	0.23	0.12	1.75E-02	1.75E-02	*
	RaTVA	0.04	0.03	7.67E-01	6.08E-01	ns

Supplementary Table S11: Comparing the performance of the RSA models. This table reports the significance of the RSA brain-model difference. We compared the correlation coefficients between brain RDM and VLS RDM with those from the brain RDM and LIN RDM within participants and hemispheres using one-tailed tests, based on the a priori hypothesis that the VLS models would exhibit greater brain-model correlations than the LIN models (cf. Methods).

Model	ROI	Accuracy (%)	s.e.m.	T	dof	p-val	unc.	sig.	CI95%	cohen-d	BF10	power
LIN	LA1	43.33 ± 2.22	0.51	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RA1	50.83 ± 1.98	0.46	1.83E+00	19	4.14E-02	*		[50.05, inf]	4.10E-01	1.89E+00	0.55
	LmTVA	38.89 ± 0.00	0.00	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RmTVA	61.39 ± 1.21	0.28	4.10E+01	19	2.61E-20	****		[60.91, inf]	9.17E+00	1.04E+17	1.00
	LpTVA	66.67 ± 0.02	0.00	4.59E+03	19	3.32E-59	****		[66.66, inf]	1.03E+03	6.88E+33	1.00
	RpTVA	77.50 ± 1.21	0.28	9.90E+01	19	1.51E-27	****		[77.02, inf]	2.21E+01	7.38E+23	1.00
	LaTVA	44.44 ± 0.00	0.00	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RaTVA	44.44 ± 0.00	0.00	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	A1	47.08 ± 4.30	0.69	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	mTVA	50.14 ± 11.28	1.81	8.00E-02	39	4.70E-01	ns		[47.09, inf]	1.00E-02	3.42E-01	0.06
	pTVA	72.08 ± 5.48	0.88	2.51E+01	39	5.18E-26	****		[70.60, inf]	3.98E+00	5.89E+22	1.00
	aTVA	44.44 ± 0.00	0.00	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	TVAs	55.56 ± 13.94	1.28	4.35E+00	119	1.47E-05	****		[53.44, inf]	4.00E-01	1.08E+03	1.00
VLS	LA1	61.94 ± 1.98	0.46	2.62E+01	19	1.08E-16	****		[61.16, inf]	5.87E+00	3.93E+13	1.00
	RA1	60.28 ± 1.98	0.46	2.26E+01	19	1.73E-15	****		[59.49, inf]	5.05E+00	2.88E+12	1.00
	LmTVA	55.56 ± 0.02	0.00	1.53E+03	19	3.86E-50	****		[55.55, inf]	3.42E+02	4.95E+33	1.00
	RmTVA	44.44 ± 0.00	0.00	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LpTVA	66.67 ± 0.02	0.00	4.59E+03	19	3.32E-59	****		[66.66, inf]	1.03E+03	6.88E+33	1.00
	RpTVA	61.11 ± 0.02	0.00	3.06E+03	19	7.36E-56	****		[61.10, inf]	6.85E+02	6.53E+33	1.00
	LaTVA	50.83 ± 1.98	0.46	1.83E+00	19	4.14E-02	*		[50.05, inf]	4.10E-01	1.89E+00	0.55
	RaTVA	44.17 ± 1.21	0.28	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	A1	61.11 ± 2.15	0.34	3.22E+01	39	4.92E-30	****		[60.53, inf]	5.10E+00	4.78E+26	1.00
	mTVA	50.00 ± 5.56	0.89	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	pTVA	63.89 ± 2.78	0.44	3.12E+01	39	1.65E-29	****		[63.14, inf]	4.94E+00	1.47E+26	1.00
	aTVA	47.50 ± 3.72	0.60	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	TVAs	53.80 ± 8.33	0.76	4.97E+00	119	1.14E-06	****		[52.53, inf]	4.50E-01	1.20E+04	1.00

Supplementary Table S12: Assessing the significance of speaker gender decoding performance using VLS and LIN models based on voxel activity. This table reports the significance of the speaker's gender decoding performance. Linear classifiers were pre-trained to detect speaker gender (2 classes) from either the VLS or the LIN models. The speaker gender of the 18 Test Stimuli (3 participants x 6 stimuli per participant) was classified using either the VLS coordinates, or the LIN features with these classifiers. We used one-sample t-tests to compare the mean of the accuracy distribution across 20 random classifier initializations (20 classifiers trained with a different initialization seed) with a chance level of 50%. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each model and ROI.

Model	ROI	Accuracy (%)	s.e.m.	T	dof	p-val	unc.	sig.	CI95%	cohen-d	BF10	power
LIN	LA1	50.42 ± 4.15	0.95	4.40E-01	19	3.33E-01	ns		[48.77, inf]	1.00E-01	5.07E-01	0.11
	RA1	10.83 ± 3.82	0.88	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LmTVA	44.17 ± 3.82	0.88	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RmTVA	50.42 ± 6.71	1.54	2.70E-01	19	3.95E-01	ns		[47.76, inf]	6.00E-02	4.80E-01	0.08
	LpTVA	52.50 ± 3.82	0.88	2.85E+00	19	5.08E-03	**		[50.99, inf]	6.40E-01	1.01E+01	0.87
	RpTVA	56.67 ± 3.33	0.76	8.72E+00	19	2.29E-08	****		[55.34, inf]	1.95E+00	6.13E+05	1.00
	LaTVA	52.50 ± 3.82	0.88	2.85E+00	19	5.08E-03	**		[50.99, inf]	6.40E-01	1.01E+01	0.87
	RaTVA	75.42 ± 6.17	1.41	1.80E+01	19	1.11E-13	****		[72.97, inf]	4.02E+00	5.71E+10	1.00
	A1	30.62 ± 20.19	3.23	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	mTVA	47.29 ± 6.29	1.01	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	pTVA	54.58 ± 4.15	0.66	6.90E+00	39	1.45E-08	****		[53.46, inf]	1.09E+00	9.41E+05	1.00
	aTVA	63.96 ± 12.55	2.01	6.94E+00	39	1.28E-08	****		[60.57, inf]	1.10E+00	1.06E+06	1.00
	TVAs	55.28 ± 10.86	1.00	5.30E+00	119	2.69E-07	****		[53.63, inf]	4.80E-01	4.69E+04	1.00
VLS	LA1	66.67 ± 0.02	0.00	4.59E+03	19	3.32E-59	****		[66.66, inf]	1.03E+03	6.88E+33	1.00
	RA1	8.33 ± 0.00	0.00	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LmTVA	49.17 ± 12.61	2.89	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RmTVA	41.67 ± 0.00	0.00	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LpTVA	58.33 ± 0.02	0.00	2.30E+03	19	1.74E-53	****		[58.33, inf]	5.14E+02	6.07E+33	1.00
	RpTVA	71.67 ± 4.08	0.94	2.31E+01	19	1.11E-15	****		[70.05, inf]	5.17E+00	4.36E+12	1.00
	LaTVA	56.67 ± 3.33	0.76	8.72E+00	19	2.29E-08	****		[55.34, inf]	1.95E+00	6.13E+05	1.00
	RaTVA	64.17 ± 3.82	0.88	1.62E+01	19	7.29E-13	****		[62.65, inf]	3.62E+00	9.69E+09	1.00
	A1	37.50 ± 29.17	4.67	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	mTVA	45.42 ± 9.67	1.55	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	pTVA	65.00 ± 7.26	1.16	1.29E+01	39	6.05E-16	****		[63.04, inf]	2.04E+00	1.05E+13	1.00
	aTVA	60.42 ± 5.19	0.83	1.25E+01	39	1.46E-15	****		[59.02, inf]	1.98E+00	4.51E+12	1.00
	TVAs	56.94 ± 11.30	1.04	6.70E+00	119	3.59E-10	****		[55.23, inf]	6.10E-01	2.64E+07	1.00

Supplementary Table S13: Assessing the significance of speaker age decoding performance using VLS and LIN models based on voxel activity. This table reports the significance of the speaker age decoding performance. Linear classifiers were pre-trained to detect speaker age (2 classes) from either the VLS or the LIN models. The speaker age of the 18 Test Stimuli (3 participants x 6 stimuli per participant) was classified using either the VLS or LIN coordinates with these classifiers. We used one-sample t-tests to compare the mean of the accuracy distribution across 20 random classifier initializations (20 classifiers trained with a different initialization seed) with a chance level of 50%. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each model and ROI.

Model	ROI	Accuracy (%)	s.e.m.	T	dof	p-val	unc.	sig.	CI95%	cohen-d	BF10	power
LIN	LA1	0.29 ± 1.28	0.29	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RA1	18.09 ± 3.26	0.75	1.63E+01	19	6.14E-13	****		[16.80, inf]	3.65E+00	1.14E+10	1.00
	LmTVA	11.18 ± 4.01	0.92	5.75E+00	19	7.61E-06	****		[9.59, inf]	1.29E+00	3.01E+03	1.00
	RmTVA	2.35 ± 3.03	0.69	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LpTVA	12.21 ± 3.39	0.78	8.13E+00	19	6.54E-08	****		[10.86, inf]	1.82E+00	2.32E+05	1.00
	RpTVA	6.76 ± 4.66	1.07	8.30E-01	19	2.10E-01	ns		[4.92, inf]	1.80E-01	6.29E-01	0.20
	LaTVA	11.47 ± 1.28	0.29	1.90E+01	19	4.04E-14	****		[10.96, inf]	4.25E+00	1.48E+11	1.00
	RaTVA	7.35 ± 8.29	1.90	7.70E-01	19	2.25E-01	ns		[4.06, inf]	1.70E-01	6.07E-01	0.18
	A1	9.19 ± 9.24	1.48	2.24E+00	39	1.55E-02	*		[6.70, inf]	3.50E-01	3.15E+00	0.71
	mTVA	6.76 ± 5.67	0.91	9.70E-01	39	1.68E-01	ns		[5.24, inf]	1.50E-01	5.30E-01	0.25
	pTVA	9.49 ± 4.90	0.78	4.59E+00	39	2.24E-05	****		[8.16, inf]	7.30E-01	1.01E+03	1.00
	aTVA	9.41 ± 6.28	1.01	3.51E+00	39	5.75E-04	***		[7.72, inf]	5.50E-01	5.39E+01	0.96
	TVAs	8.55 ± 5.78	0.53	5.04E+00	119	8.44E-07	****		[7.68, inf]	4.60E-01	1.59E+04	1.00
VLS	LA1	0.15 ± 0.64	0.15	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RA1	11.47 ± 5.09	1.17	4.79E+00	19	6.37E-05	****		[9.45, inf]	1.07E+00	4.49E+02	1.00
	LmTVA	11.47 ± 4.73	1.09	5.15E+00	19	2.87E-05	****		[9.59, inf]	1.15E+00	9.13E+02	1.00
	RmTVA	0.59 ± 1.50	0.34	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LpTVA	9.71 ± 3.37	0.77	4.95E+00	19	4.44E-05	****		[8.37, inf]	1.11E+00	6.19E+02	1.00
	RpTVA	22.65 ± 2.10	0.48	3.48E+01	19	5.69E-19	****		[21.81, inf]	7.78E+00	5.62E+15	1.00
	LaTVA	10.29 ± 4.51	1.03	4.27E+00	19	2.09E-04	***		[8.51, inf]	9.50E-01	1.57E+02	0.99
	RaTVA	6.18 ± 3.94	0.90	3.30E-01	19	3.74E-01	ns		[4.62, inf]	7.00E-02	4.88E-01	0.09
	A1	5.81 ± 6.72	1.08	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	mTVA	6.03 ± 6.48	1.04	1.40E-01	39	4.44E-01	ns		[4.28, inf]	2.00E-02	3.44E-01	0.07
	pTVA	16.18 ± 7.05	1.13	9.12E+00	39	1.65E-11	****		[14.27, inf]	1.44E+00	5.85E+08	1.00
	aTVA	8.24 ± 4.71	0.75	3.12E+00	39	1.69E-03	**		[6.97, inf]	4.90E-01	2.09E+01	0.92
	TVAs	10.15 ± 7.55	0.69	6.17E+00	119	4.96E-09	****		[9.00, inf]	5.60E-01	2.12E+06	1.00

Supplementary Table S14. Assessing the significance of speaker identity decoding performance using VLS and LIN models based on voxel activity. This table reports the significance of the speaker age decoding performance. Linear classifiers were pre-trained to detect speaker age (2 classes) from either the VLS or the LIN models. Linear classifiers were pre-trained to detect speaker identity (17 classes) from either the VLS or the LIN models. The speaker identity of the 18 Test Stimuli (3 participants x 6 stimuli per participant) was classified using either the VLS or LIN coordinates with these classifiers. We used one-sample t-tests to compare the mean of the accuracy distribution across 20 random classifier initializations (20 classifiers trained with a different initialization seed) with a chance level of 5.88%. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each model and ROI.

Category	ROI	Accuracy VLS (%)	Accuracy LIN (%)	s.e.m. VLS	s.e.m. LIN	T VLS vs LIN	dof	p-val	unc.	sig.	CI95%	cohen-d	BF10	power
Gender	LA1	61.94 ± 1.98	43.33 ± 2.22	0.46	0.51	3.06E+01	19	1.24E-17	****		[17.34, 19.88]	8.610000	2.94E+14	1.00
	RA1	60.28 ± 1.98	50.83 ± 1.98	0.46	0.46	1.62E+01	19	1.46E-12	****		[8.22, 10.67]	4.640000	4.85E+09	1.00
	LmTVA	55.56 ± 0.02	38.89 ± 0.02	0.00	0.00	INF	19	0.00E+00	****		[nan, nan]	1027.400000	nan	1.00
	RmTVA	44.44 ± 0.02	61.39 ± 1.21	0.00	0.28	-6.10E+01	19	2.91E-23	****		[-17.53, -16.36]	19.290000	6.23E+19	1.00
	LpTVA	66.67 ± 0.02	66.67 ± 0.02	0.00	0.00	nan	19	nan	ns		[nan, nan]	0.000000	nan	0.05
	RpTVA	61.11 ± 0.02	77.50 ± 1.21	0.00	0.28	-5.90E+01	19	5.47E-23	****		[-16.97, -15.81]	18.660000	3.43E+19	1.00
	LaTVA	50.83 ± 1.98	44.44 ± 0.02	0.46	0.00	1.41E+01	19	1.65E-11	****		[5.44, 7.34]	4.440000	4.96E+08	1.00
	RaTVA	44.17 ± 1.21	44.44 ± 0.02	0.28	0.00	-1.00E+00	19	3.30E-01	ns		[-0.86, 0.30]	0.320000	3.61E-01	0.27
	A1	61.11 ± 2.15	47.08 ± 4.30	0.34	0.69	1.66E+01	39	2.67E-19	****		[12.32, 15.73]	4.070000	1.78E+16	1.00
	mTVA	50.00 ± 5.56	50.14 ± 11.28	0.89	1.81	-5.00E-02	39	9.59E-01	ns		[-5.59, 5.31]	0.020000	1.71E-01	0.05
	pTVA	63.89 ± 2.78	72.08 ± 5.48	0.44	0.88	-6.21E+00	39	2.64E-07	****		[-10.86, -5.53]	1.860000	5.92E+04	1.00
	aTVA	47.50 ± 3.72	44.44 ± 0.01	0.60	0.00	5.14E+00	39	8.05E-06	****		[1.85, 4.26]	1.150000	2.45E+03	1.00
	TVAs	53.80 ± 8.33	55.56 ± 13.94	0.76	1.28	-1.60E+00	119	1.12E-01	ns		[-3.94, 0.42]	0.150000	3.49E-01	0.38
Age	LA1	66.67 ± 0.02	50.42 ± 4.15	0.00	0.95	1.71E+01	19	5.58E-13	****		[14.26, 18.24]	5.400000	1.20E+10	1.00
	RA1	8.33 ± 0.02	10.83 ± 3.82	0.00	0.88	-2.85E+00	19	1.03E-02	*		[-4.34, -0.66]	0.900000	5.02E+00	0.97
	LmTVA	49.17 ± 12.61	44.17 ± 3.82	2.89	0.88	1.71E+00	19	1.04E-01	ns		[-1.12, 11.12]	0.520000	7.97E-01	0.60
	RmTVA	41.67 ± 0.02	50.42 ± 6.71	0.00	1.54	-5.69E+00	19	1.76E-05	****		[-11.97, -5.53]	1.800000	1.32E+03	1.00
	LpTVA	58.33 ± 0.02	52.50 ± 3.82	0.00	0.88	6.67E+00	19	2.24E-06	****		[4.00, 7.66]	2.110000	8.58E+03	1.00
	RpTVA	71.67 ± 4.08	56.67 ± 3.33	0.94	0.76	1.16E+01	19	4.80E-10	****		[12.29, 17.71]	3.920000	2.12E+07	1.00
	LaTVA	56.67 ± 3.33	52.50 ± 3.82	0.76	0.88	3.68E+00	19	1.58E-03	**		[1.80, 6.53]	1.130000	2.46E+01	1.00
	RaTVA	64.17 ± 3.82	75.42 ± 6.17	0.88	1.41	-6.90E+00	19	1.40E-06	****		[-14.66, -7.84]	2.140000	1.31E+04	1.00
	A1	37.50 ± 29.17	30.62 ± 20.19	4.67	3.23	4.21E+00	39	1.43E-04	***		[3.58, 10.17]	0.270000	1.75E+02	0.39
	mTVA	45.42 ± 9.67	47.29 ± 6.29	1.55	1.01	-9.50E-01	39	3.47E-01	ns		[-5.86, 2.11]	0.230000	2.60E-01	0.29
	pTVA	65.00 ± 7.26	54.58 ± 4.15	1.16	0.66	9.78E+00	39	4.83E-12	****		[8.26, 12.57]	1.740000	1.83E+09	1.00
	aTVA	60.42 ± 5.19	63.96 ± 12.55	0.83	2.01	-2.25E+00	39	3.03E-02	*		[-6.73, -0.35]	0.360000	1.61E+00	0.61
	TVAs	56.94 ± 11.30	55.28 ± 10.86	1.04	1.00	1.56E+00	119	1.22E-01	ns		[-0.45, 3.78]	0.150000	3.28E-01	0.37
Identity	LA1	0.15 ± 0.64	0.29 ± 1.28	0.15	0.29	-4.40E-01	19	6.66E-01	ns		[-0.85, 0.56]	0.140000	2.53E-01	0.09
	RA1	11.47 ± 5.09	18.09 ± 3.26	1.17	0.75	-4.58E+00	19	2.05E-04	***		[-9.64, -3.59]	1.510000	1.47E+02	1.00
	LmTVA	11.47 ± 4.73	11.18 ± 4.01	1.09	0.92	2.10E-01	19	8.39E-01	ns		[-2.70, 3.29]	0.070000	2.37E-01	0.06
	RmTVA	0.59 ± 1.50	2.35 ± 3.03	0.34	0.69	-2.11E+00	19	4.86E-02	*		[-3.52, -0.01]	0.720000	1.42E+00	0.86
	LpTVA	9.71 ± 3.37	12.21 ± 3.39	0.77	0.78	-2.43E+00	19	2.53E-02	*		[-4.65, -0.35]	0.720000	2.39E+00	0.86
	RpTVA	22.65 ± 2.10	6.76 ± 4.66	0.48	1.07	1.31E+01	19	5.99E-11	****		[13.34, 18.42]	4.280000	1.48E+08	1.00
	LaTVA	10.29 ± 4.51	11.47 ± 1.28	1.03	0.29	-1.00E+00	19	3.30E-01	ns		[-3.64, 1.29]	0.350000	3.61E-01	0.31
	RaTVA	6.18 ± 3.94	7.35 ± 8.29	0.90	1.90	-7.50E-01	19	4.64E-01	ns		[-4.47, 2.12]	0.180000	2.98E-01	0.12
	A1	5.81 ± 6.72	9.19 ± 9.24	1.08	1.48	-3.77E+00	39	5.40E-04	***		[-5.20, -1.57]	0.410000	5.30E+01	0.72
	mTVA	6.03 ± 6.48	6.76 ± 5.67	1.04	0.91	-8.80E-01	39	3.83E-01	ns		[-2.42, 0.95]	0.120000	2.45E-01	0.11
	pTVA	16.18 ± 7.05	9.49 ± 4.90	1.13	0.78	4.01E+00	39	2.65E-04	***		[3.32, 10.07]	1.090000	1.00E+02	1.00
	aTVA	8.24 ± 4.71	9.41 ± 6.28	0.75	1.01	-1.21E+00	39	2.32E-01	ns		[-3.14, 0.79]	0.210000	3.37E-01	0.25
	TVAs	10.15 ± 7.55	8.55 ± 5.78	0.69	0.53	2.07E+00	119	4.06E-02	*		[0.07, 3.12]	0.240000	7.94E-01	0.73

Supplementary Table S15: Comparing the performance of the models decoding speaker identity-related information. This table reports the significance of the speaker identity decoding VLS-LIN difference. Paired t-tests were conducted between the mean scores of linear classifiers pre-trained to detect gender (2 classes), age (2 classes), and identity (17 classes) from the VLS features and those trained with the LIN features. These scores were obtained after classifying the VLS or LIN coordinates of the 18 Test Stimuli (3 participants x 6 stimuli per participant). s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each speaker information and ROI.

Category	Model	ROI	Accuracy ROI (%)	Accuracy A1 (%)	s.e.m. ROI	s.e.m. A1	T ROI vs A1	dof	p-val	unc. sig.	CI95%	cohen-d	BF10	power
Gender	LIN	mTVA	50.14 ± 11.28	47.08 ± 4.30	1.81	0.69	1.58E+00	78	1.20E-01	ns	[-0.79, 6.90]	3.50E-01	6.83E-01	0.35
		pTVA	72.08 ± 5.48	47.08 ± 4.30	0.88	0.69	2.24E+01	78	0.00E+00	****	[22.78, 27.22]	5.01E+00	1.30E+32	1.00
		aTVA	44.44 ± 0.00	47.08 ± 4.30	0.00	0.69	-3.83E+00	78	0.00E+00	***	[-4.01, -1.27]	8.60E-01	9.78E+01	0.97
		TVAs	55.56 ± 13.94	47.08 ± 4.30	1.28	0.69	3.76E+00	158	0.00E+00	***	[4.02, 12.92]	6.90E-01	9.90E+01	0.96
	VLS	mTVA	50.00 ± 5.56	61.11 ± 2.15	0.89	0.34	-1.17E+01	78	0.00E+00	****	[-13.01, -9.21]	2.60E+00	3.45E+15	1.00
		pTVA	63.89 ± 2.78	61.11 ± 2.15	0.44	0.34	4.94E+00	78	0.00E+00	****	[1.66, 3.90]	1.10E+00	3.59E+03	1.00
		aTVA	47.50 ± 3.72	61.11 ± 2.15	0.60	0.34	-1.98E+01	78	0.00E+00	****	[-14.98, -12.24]	4.43E+00	4.06E+28	1.00
		TVAs	53.80 ± 8.33	61.11 ± 2.15	0.76	0.34	-5.46E+00	158	0.00E+00	****	[-9.96, -4.67]	1.00E+00	6.55E+04	1.00
Age	LIN	mTVA	47.29 ± 6.29	30.62 ± 20.19	1.01	3.23	4.92E+00	78	0.00E+00	****	[9.93, 23.41]	1.10E+00	3.41E+03	1.00
		pTVA	54.58 ± 4.15	30.62 ± 20.19	0.66	3.23	7.26E+00	78	0.00E+00	****	[17.39, 30.53]	1.62E+00	3.02E+07	1.00
		aTVA	63.96 ± 12.55	30.62 ± 20.19	2.01	3.23	8.76E+00	78	0.00E+00	****	[25.75, 40.91]	1.96E+00	1.68E+10	1.00
		TVAs	55.28 ± 10.86	30.62 ± 20.19	1.00	3.23	9.72E+00	158	0.00E+00	****	[19.65, 29.66]	1.78E+00	4.55E+14	1.00
	VLS	mTVA	45.42 ± 9.67	37.50 ± 29.17	1.55	4.67	1.61E+00	78	1.10E-01	ns	[-1.88, 17.71]	3.60E-01	7.10E-01	0.36
		pTVA	65.00 ± 7.26	37.50 ± 29.17	1.16	4.67	5.71E+00	78	0.00E+00	****	[17.92, 37.08]	1.28E+00	6.21E+04	1.00
		aTVA	60.42 ± 5.19	37.50 ± 29.17	0.83	4.67	4.83E+00	78	0.00E+00	****	[13.47, 32.36]	1.08E+00	2.48E+03	1.00
		TVAs	56.94 ± 11.30	37.50 ± 29.17	1.04	4.67	6.03E+00	158	0.00E+00	****	[13.07, 25.82]	1.10E+00	8.68E+05	1.00
Identity	LIN	mTVA	6.76 ± 5.67	9.19 ± 9.24	0.91	1.48	-1.40E+00	78	1.70E-01	ns	[-5.88, 1.03]	3.10E-01	5.42E-01	0.28
		pTVA	9.49 ± 4.90	9.19 ± 9.24	0.78	1.48	1.80E-01	78	8.60E-01	ns	[-3.04, 3.63]	4.00E-02	2.36E-01	0.05
		aTVA	9.41 ± 6.28	9.19 ± 9.24	1.01	1.48	1.20E-01	78	9.00E-01	ns	[-3.34, 3.78]	3.00E-02	2.34E-01	0.05
		TVAs	8.55 ± 5.78	9.19 ± 9.24	0.53	1.48	-5.10E-01	158	6.10E-01	ns	[-3.11, 1.83]	9.00E-02	2.19E-01	0.08
	VLS	mTVA	6.03 ± 6.48	5.81 ± 6.72	1.04	1.08	1.50E-01	78	8.80E-01	ns	[-2.76, 3.20]	3.00E-02	2.35E-01	0.05
		pTVA	16.18 ± 7.05	5.81 ± 6.72	1.13	1.08	6.65E+00	78	0.00E+00	****	[7.26, 13.47]	1.49E+00	2.43E+06	1.00
		aTVA	8.24 ± 4.71	5.81 ± 6.72	0.75	1.08	1.85E+00	78	7.00E-02	ns	[-0.19, 5.04]	4.10E-01	1.01E+00	0.45
		TVAs	10.15 ± 7.55	5.81 ± 6.72	0.69	1.08	3.21E+00	158	0.00E+00	**	[1.67, 7.00]	5.90E-01	1.91E+01	0.89

Supplementary Table S16: Comparing the performance of the models decoding speaker identity-related information by ROI. This table reports the significance of the speaker identity decoding A1-TVAs difference. Two-sample t-tests were conducted for each model to determine if there was an A1-TVAs difference between the mean scores of linear classifiers pre-trained to detect gender (2 classes), age (2 classes), and identity (17 classes). These scores were obtained by classifying the VLS coordinates or LIN features, reconstructed by different ROIs, for the 18 Test Stimuli (3 participants x 6 stimuli per participant). s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each speaker information and model.

Model	ROI	Accuracy (%)	s.e.m.	T	dof	p-val	unc.	sig.	CI95%	cohen-d	BF10	power
LIN	LA1	44.87 ± 7.99	2.31	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RA1	51.28 ± 10.02	2.89	4.40E-01	12	3.33E-01	ns		[46.13, inf]	1.20E-01	6.06E-01	0.11
	LmTVA	51.71 ± 10.76	3.11	5.50E-01	12	2.96E-01	ns		[46.17, inf]	1.50E-01	6.35E-01	0.13
	RmTVA	43.59 ± 8.40	2.42	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LpTVA	50.00 ± 8.72	2.52	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RpTVA	52.99 ± 10.36	2.99	1.00E+00	12	1.69E-01	ns		[47.66, inf]	2.80E-01	8.49E-01	0.24
	LaTVA	51.28 ± 8.48	2.45	5.20E-01	12	3.05E-01	ns		[46.92, inf]	1.50E-01	6.27E-01	0.13
	RaTVA	45.30 ± 9.71	2.80	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	A1	48.08 ± 9.62	1.92	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	mTVA	47.65 ± 10.47	2.09	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	pTVA	51.50 ± 9.69	1.94	7.70E-01	25	2.24E-01	ns		[48.18, inf]	1.50E-01	5.43E-01	0.19
	aTVA	48.29 ± 9.59	1.92	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	TVAs	49.15 ± 10.07	1.15	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
VLS	LA1	50.00 ± 11.32	3.27	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	RA1	61.54 ± 6.34	1.83	6.31E+00	12	1.96E-05	****		[58.28, inf]	1.75E+00	1.31E+03	1.00
	LmTVA	51.71 ± 5.06	1.46	1.17E+00	12	1.32E-01	ns		[49.11, inf]	3.20E-01	9.85E-01	0.29
	RmTVA	45.73 ± 8.76	2.53	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	LpTVA	63.25 ± 7.40	2.14	6.20E+00	12	2.29E-05	****		[59.44, inf]	1.72E+00	1.14E+03	1.00
	RpTVA	60.26 ± 6.48	1.87	5.48E+00	12	7.01E-05	****		[56.92, inf]	1.52E+00	4.30E+02	1.00
	LaTVA	60.26 ± 6.10	1.76	5.82E+00	12	4.10E-05	****		[57.12, inf]	1.61E+00	6.86E+02	1.00
	RaTVA	50.00 ± 8.98	2.59	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	A1	55.77 ± 10.84	2.17	2.66E+00	25	6.70E-03	**		[52.07, inf]	5.20E-01	7.39E+00	0.83
	mTVA	48.72 ± 7.75	1.55	n/a	n/a	n/a	n/a		n/a	n/a	n/a	n/a
	pTVA	61.75 ± 7.12	1.42	8.26E+00	25	6.56E-09	****		[59.32, inf]	1.62E+00	2.00E+06	1.00
	aTVA	55.13 ± 9.24	1.85	2.78E+00	25	5.13E-03	**		[51.97, inf]	5.40E-01	9.24E+00	0.85
	TVAs	55.20 ± 9.68	1.10	4.71E+00	77	5.29E-06	****		[53.36, inf]	5.30E-01	3.23E+03	1.00

Supplementary Table S17: Assessing the significance of the speaker gender categorization task. This table reports the significance of the speaker's gender categorization performance. 342 voice stimuli were used in the experiments: the original stimuli (N = 18), directly reconstructed stimuli using the LIN and the VLS models (N = 36), and brain-reconstructed stimuli (18 stimuli x 2 models x 4 regions of interest x 2 hemispheres, N = 288). The participants were tasked with identifying the gender of the presented voice in each trial by clicking either the 'Female' or 'Male' button. To evaluate the accuracy of the binary responses, we computed the classification accuracy for each participant and region of interest (ROI). We then utilized one-sample t-tests to compare the mean accuracy distribution across all participants to the chance level of 50%. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each model and ROI.

Model	ROI	Accuracy (%)	s.e.m.	T	dof	p-val	unc. sig.	CI95%	cohen-d	BF10	power
LIN	LA1	46.15 ± 14.48	4.18	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RA1	44.23 ± 11.50	3.32	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	LmTVA	50.00 ± 9.81	2.83	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RmTVA	57.69 ± 12.85	3.71	2.07E+00	12	3.02E-02	*	[51.08, inf]	5.80E-01	2.80E+00	0.62
	LpTVA	50.00 ± 10.34	2.98	0.00E+00	12	5.00E-01	ns	[44.68, inf]	0.00E+00	5.56E-01	0.05
	RpTVA	50.64 ± 10.57	3.05	2.10E-01	12	4.19E-01	ns	[45.20, inf]	6.00E-02	5.67E-01	0.07
	LaTVA	48.72 ± 13.01	3.76	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RaTVA	62.82 ± 13.32	3.85	3.33E+00	12	2.98E-03	**	[55.97, inf]	9.20E-01	1.77E+01	0.93
	A1	45.19 ± 13.11	2.62	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	mTVA	53.85 ± 12.06	2.41	1.59E+00	25	6.17E-02	ns	[49.73, inf]	3.10E-01	1.26E+00	0.46
	pTVA	50.32 ± 10.46	2.09	1.50E-01	25	4.40E-01	ns	[46.75, inf]	3.00E-02	4.19E-01	0.07
	aTVA	55.77 ± 14.94	2.99	1.93E+00	25	3.24E-02	*	[50.67, inf]	3.80E-01	2.06E+00	0.59
	TVAs	53.31 ± 12.82	1.46	2.27E+00	77	1.31E-02	*	[50.88, inf]	2.60E-01	2.77E+00	0.73
VLS	LA1	54.49 ± 10.65	3.07	1.46E+00	12	8.50E-02	ns	[49.01, inf]	4.00E-01	1.32E+00	0.39
	RA1	50.00 ± 8.01	2.31	0.00E+00	12	5.00E-01	ns	[45.88, inf]	0.00E+00	5.56E-01	0.05
	LmTVA	51.28 ± 10.26	2.96	4.30E-01	12	3.36E-01	ns	[46.01, inf]	1.20E-01	6.04E-01	0.11
	RmTVA	54.49 ± 10.65	3.07	1.46E+00	12	8.50E-02	ns	[49.01, inf]	4.00E-01	1.32E+00	0.39
	LpTVA	45.51 ± 11.14	3.22	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RpTVA	56.41 ± 8.74	2.52	2.54E+00	12	1.30E-02	*	[51.91, inf]	7.00E-01	5.38E+00	0.77
	LaTVA	64.74 ± 7.42	2.14	6.88E+00	12	8.46E-06	****	[60.93, inf]	1.91E+00	2.74E+03	1.00
	RaTVA	61.54 ± 14.81	4.28	2.70E+00	12	9.68E-03	**	[53.92, inf]	7.50E-01	6.79E+00	0.82
	A1	52.24 ± 9.68	1.94	1.16E+00	25	1.29E-01	ns	[48.94, inf]	2.30E-01	7.57E-01	0.30
	mTVA	52.88 ± 10.58	2.12	1.36E+00	25	9.24E-02	ns	[49.27, inf]	2.70E-01	9.47E-01	0.38
	pTVA	50.96 ± 11.40	2.28	4.20E-01	25	3.38E-01	ns	[47.07, inf]	8.00E-02	4.50E-01	0.11
	aTVA	63.14 ± 11.82	2.36	5.56E+00	25	4.45E-06	****	[59.10, inf]	1.09E+00	4.79E+03	1.00
	TVAs	55.66 ± 12.48	1.42	3.98E+00	77	7.72E-05	****	[53.29, inf]	4.50E-01	2.69E+02	0.99

Supplementary Table S18: Assessing the significance of the speaker age categorization task. This table reports the significance of the speaker age categorization performance. 342 voice stimuli were used in the experiments: the original stimuli (N = 18), directly reconstructed stimuli using the LIN and the VLS models (N = 36), and brain-reconstructed stimuli (18 stimuli x 2 models x 4 regions of interest x 2 hemispheres, N = 288). The participants were tasked with identifying the approximate age of the presented voice in each trial by clicking either the 'Younger' or 'Older' button. To evaluate the accuracy of the binary responses, we computed the classification accuracy for each participant and region of interest (ROI). We then utilized one-sample t-tests to compare the mean accuracy distribution across all participants to the chance level of 50%. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each model and ROI.

Model	ROI	Accuracy (%)	s.e.m.	T	dof	p-val	unc. sig.	CI95%	cohen-d	BF10	power
LIN	LA1	54.70 ± 9.89	3.50	1.34E+00	8	1.08E-01	ns	[48.20, inf]	4.50E-01	1.29E+00	0.34
	RA1	57.41 ± 8.69	3.07	2.41E+00	8	2.12E-02	*	[51.70, inf]	8.00E-01	4.14E+00	0.71
	LmTVA	57.04 ± 7.77	2.75	2.56E+00	8	1.68E-02	*	[51.93, inf]	8.50E-01	4.94E+00	0.76
	RmTVA	42.36 ± 8.22	2.91	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	LpTVA	57.78 ± 7.70	2.72	2.86E+00	8	1.06E-02	*	[52.72, inf]	9.50E-01	7.04E+00	0.83
	RpTVA	50.98 ± 9.61	3.40	2.90E-01	8	3.90E-01	ns	[44.67, inf]	1.00E-01	6.67E-01	0.08
	LaTVA	40.48 ± 9.52	3.37	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RaTVA	40.52 ± 7.04	2.49	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	A1	56.05 ± 9.41	2.28	2.65E+00	17	8.36E-03	**	[52.09, inf]	6.30E-01	6.92E+00	0.82
	mTVA	49.70 ± 10.85	2.63	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	pTVA	54.38 ± 9.34	2.27	1.93E+00	17	3.51E-02	*	[50.44, inf]	4.60E-01	2.22E+00	0.58
	aTVA	40.50 ± 8.37	2.03	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	TVAs	48.19 ± 11.18	1.54	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
VLS	LA1	72.22 ± 9.16	3.24	6.86E+00	8	6.48E-05	****	[66.20, inf]	2.29E+00	4.52E+02	1.00
	RA1	48.33 ± 5.77	2.04	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	LmTVA	51.46 ± 7.76	2.74	5.30E-01	8	3.04E-01	ns	[46.36, inf]	1.80E-01	7.25E-01	0.12
	RmTVA	41.11 ± 6.57	2.32	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	LpTVA	60.61 ± 5.67	2.00	5.29E+00	8	3.68E-04	***	[56.88, inf]	1.76E+00	1.06E+02	1.00
	RpTVA	66.05 ± 6.65	2.35	6.83E+00	8	6.70E-05	****	[61.68, inf]	2.28E+00	4.40E+02	1.00
	LaTVA	52.02 ± 8.33	2.94	6.90E-01	8	2.56E-01	ns	[46.54, inf]	2.30E-01	7.83E-01	0.15
	RaTVA	50.00 ± 7.53	2.66	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	A1	60.28 ± 14.19	3.44	2.99E+00	17	4.14E-03	**	[54.29, inf]	7.00E-01	1.24E+01	0.89
	mTVA	46.29 ± 8.86	2.15	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	pTVA	63.33 ± 6.75	1.64	8.14E+00	17	1.44E-07	****	[60.48, inf]	1.92E+00	1.11E+05	1.00
	aTVA	51.01 ± 8.00	1.94	5.20E-01	17	3.05E-01	ns	[47.63, inf]	1.20E-01	5.49E-01	0.13
	TVAs	53.54 ± 10.69	1.47	2.41E+00	53	9.69E-03	**	[51.08, inf]	3.30E-01	4.15E+00	0.77

Supplementary Table S19: Assessing the significance of the speaker identity discrimination task. This table reports the significance of the speaker identity discrimination performance. The participants listened to 684 voice stimuli with short breaks in between. Each trial contained 2 short sound samples, and the participants had to indicate whether the samples were from the same speaker or different speakers. We then utilized one-sample t-tests to compare the mean accuracy distribution across all participants to the chance level of 50%. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each model and ROI.

Category	ROI	Accuracy VLS (%)	Accuracy LIN (%)	s.e.m. VLS	s.e.m. LIN	T VLS vs LIN	dof	p-val	unc. sig.	CI95%	cohen-d	BF10	power
Gender	LA1	48.33 ± 10.56	46.11 ± 6.60	3.52	2.20	4.70E-01	9	6.48E-01	ns	[-8.41, 12.85]	0.240000	3.40E-01	0.10
	RA1	60.00 ± 5.98	53.33 ± 8.31	1.99	2.77	1.86E+00	9	9.63E-02	ns	[-1.46, 14.79]	0.870000	1.08E+00	0.69
	LmTVA	51.67 ± 5.58	50.56 ± 10.08	1.86	3.36	4.10E-01	9	6.93E-01	ns	[-5.05, 7.27]	0.130000	3.32E-01	0.07
	RmTVA	46.67 ± 9.03	43.33 ± 8.53	3.01	2.84	1.20E+00	9	2.60E-01	ns	[-2.94, 9.60]	0.360000	5.51E-01	0.18
	LpTVA	63.89 ± 7.95	49.44 ± 9.44	2.65	3.15	3.03E+00	9	1.43E-02	*	[3.65, 25.24]	1.570000	4.66E+00	0.99
	RpTVA	62.22 ± 4.84	53.33 ± 10.60	1.61	3.53	1.95E+00	9	8.26E-02	ns	[-1.41, 19.18]	1.020000	1.21E+00	0.82
	LaTVA	60.00 ± 5.44	50.56 ± 8.77	1.81	2.92	2.68E+00	9	2.50E-02	*	[1.49, 17.40]	1.230000	3.00E+00	0.93
	RaTVA	50.00 ± 9.94	45.56 ± 9.88	3.31	3.29	1.15E+00	9	2.80E-01	ns	[-4.30, 13.19]	0.430000	5.26E-01	0.23
	A1	54.17 ± 10.37	49.72 ± 8.33	2.38	1.91	1.52E+00	19	1.45E-01	ns	[-1.67, 10.56]	0.460000	6.25E-01	0.50
	mTVA	49.17 ± 7.91	46.94 ± 10.01	1.81	2.30	1.16E+00	19	2.58E-01	ns	[-1.77, 6.21]	0.240000	4.21E-01	0.18
	pTVA	63.06 ± 6.64	51.39 ± 10.23	1.52	2.35	3.57E+00	19	2.06E-03	**	[4.82, 18.51]	1.320000	1.95E+01	1.00
	aTVA	55.00 ± 9.44	48.06 ± 9.67	2.17	2.22	2.66E+00	19	1.54E-02	*	[1.49, 12.40]	0.710000	3.59E+00	0.85
	TVAs	55.74 ± 9.88	48.80 ± 10.15	1.29	1.32	4.37E+00	59	5.05E-05	****	[3.77, 10.12]	0.690000	4.08E+02	1.00
Age	LA1	54.49 ± 10.65	46.15 ± 14.48	3.07	4.18	1.54E+00	12	1.50E-01	ns	[-3.48, 20.14]	0.630000	7.18E-01	0.55
	RA1	50.00 ± 8.01	44.23 ± 11.50	2.31	3.32	1.24E+00	12	2.39E-01	ns	[-4.38, 15.92]	0.560000	5.25E-01	0.46
	LmTVA	51.28 ± 10.26	50.00 ± 9.81	2.96	2.83	2.70E-01	12	7.94E-01	ns	[-9.17, 11.73]	0.120000	2.87E-01	0.07
	RmTVA	54.49 ± 10.65	57.69 ± 12.85	3.07	3.71	-7.20E-01	12	4.88E-01	ns	[-12.97, 6.55]	0.260000	3.47E-01	0.14
	LpTVA	45.51 ± 11.14	50.00 ± 10.34	3.22	2.98	-1.17E+00	12	2.66E-01	ns	[-12.87, 3.89]	0.400000	4.90E-01	0.27
	RpTVA	56.41 ± 8.74	50.64 ± 10.57	2.52	3.05	1.74E+00	12	1.08E-01	ns	[-1.47, 13.00]	0.570000	9.09E-01	0.47
	LaTVA	64.74 ± 7.42	48.72 ± 13.01	2.14	3.76	5.25E+00	12	2.04E-04	***	[9.38, 22.68]	1.450000	1.55E+02	1.00
	RaTVA	61.54 ± 14.81	62.82 ± 13.32	4.28	3.85	-2.50E-01	12	8.08E-01	ns	[-12.51, 9.95]	0.090000	2.86E-01	0.06
	A1	52.24 ± 9.68	45.19 ± 13.11	1.94	2.62	2.01E+00	25	5.55E-02	ns	[-0.18, 14.28]	0.600000	1.16E+00	0.84
	mTVA	52.88 ± 10.58	53.85 ± 12.06	2.12	2.41	-3.00E-01	25	7.70E-01	ns	[-7.65, 5.72]	0.080000	2.16E-01	0.07
	pTVA	50.96 ± 11.40	50.32 ± 10.46	2.28	2.09	2.40E-01	25	8.14E-01	ns	[-4.90, 6.19]	0.060000	2.13E-01	0.06
	aTVA	63.14 ± 11.82	55.77 ± 14.94	2.36	2.99	2.16E+00	25	4.02E-02	*	[0.35, 14.39]	0.540000	1.50E+00	0.75
	TVAs	55.66 ± 12.48	53.31 ± 12.82	1.42	1.46	1.28E+00	77	2.03E-01	ns	[-1.29, 6.00]	0.180000	2.74E-01	0.36
Identity	LA1	72.22 ± 9.16	54.70 ± 9.89	3.24	3.50	3.64E+00	8	6.61E-03	**	[6.41, 28.63]	1.730000	8.84E+00	0.99
	RA1	48.33 ± 5.77	57.41 ± 8.69	2.04	3.07	-1.97E+00	8	8.49E-02	ns	[-19.72, 1.57]	1.160000	1.23E+00	0.86
	LmTVA	51.46 ± 7.76	57.04 ± 7.77	2.74	2.75	-1.44E+00	8	1.87E-01	ns	[-14.49, 3.34]	0.680000	7.08E-01	0.43
	RmTVA	41.11 ± 6.57	42.36 ± 8.22	2.32	2.91	-2.50E-01	8	8.08E-01	ns	[-12.72, 10.22]	0.160000	3.30E-01	0.07
	LpTVA	60.61 ± 5.67	57.78 ± 7.70	2.00	2.72	1.05E+00	8	3.23E-01	ns	[-3.37, 9.02]	0.390000	5.02E-01	0.18
	RpTVA	66.05 ± 6.65	50.98 ± 9.61	2.35	3.40	4.37E+00	8	2.39E-03	**	[7.11, 23.03]	1.720000	2.01E+01	0.99
	LaTVA	52.02 ± 8.33	40.48 ± 9.52	2.94	3.37	2.00E+00	8	8.02E-02	ns	[-1.75, 24.84]	1.220000	1.29E+00	0.89
	RaTVA	50.00 ± 7.53	40.52 ± 7.04	2.66	2.49	2.60E+00	8	3.16E-02	*	[1.07, 17.88]	1.230000	2.59E+00	0.89
	A1	60.28 ± 14.19	56.05 ± 9.41	3.44	2.28	9.20E-01	17	3.68E-01	ns	[-5.42, 13.86]	0.340000	3.54E-01	0.28
	mTVA	46.29 ± 8.86	49.70 ± 10.85	2.15	2.63	-1.10E+00	17	2.86E-01	ns	[-9.96, 3.13]	0.330000	4.11E-01	0.27
	pTVA	63.33 ± 6.75	54.38 ± 9.34	1.64	2.27	3.46E+00	17	3.02E-03	**	[3.49, 14.41]	1.070000	1.45E+01	0.99
	aTVA	51.01 ± 8.00	40.50 ± 8.37	1.94	2.03	3.17E+00	17	5.62E-03	**	[3.51, 17.51]	1.250000	8.55E+00	1.00
	TVAs	53.54 ± 10.69	48.19 ± 11.18	1.47	1.54	2.80E+00	53	7.15E-03	**	[1.51, 9.18]	0.480000	4.88E+00	0.94

Supplementary Table S20: Comparing human listeners' performance in discriminating speaker identity-related information decoded with VLS versus LIN. This table reports the significance of the VLS-LIN difference in the speaker identity categorization and discrimination performance. Paired t-tests were conducted between the scores of human listeners at discriminating the speaker gender (2 classes), age (2 classes), and identity (17 classes) of the 18 Test Stimuli reconstructed from the VLS features with those from LIN features. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each speaker identity information and ROI.

Category	Model	ROI	Accuracy ROI (%)	Accuracy A1 (%)	s.e.m. ROI	s.e.m. A1	T ROI vs A1	dof	p-val	unc. sig.	CI95%	cohen-d	BF10	power
Gender	LIN	mTVA	46.94 ± 10.01	49.72 ± 8.33	2.30	1.91	-9.30E-01	38	3.60E-01	ns	[-8.83, 3.27]	2.90E-01	4.35E-01	0.15
		pTVA	51.39 ± 10.23	49.72 ± 8.33	2.35	1.91	5.50E-01	38	5.80E-01	ns	[-4.46, 7.79]	1.70E-01	3.48E-01	0.08
		aTVA	48.06 ± 9.67	49.72 ± 8.33	2.22	1.91	-5.70E-01	38	5.70E-01	ns	[-7.59, 4.26]	1.80E-01	3.51E-01	0.09
		TVAs	48.80 ± 10.15	49.72 ± 8.33	1.32	1.91	-3.60E-01	78	7.20E-01	ns	[-5.99, 4.14]	9.00E-02	2.77E-01	0.06
	VLS	mTVA	49.17 ± 7.91	54.17 ± 10.37	1.81	2.38	-1.67E+00	38	1.00E-01	ns	[-11.06, 1.06]	5.30E-01	9.19E-01	0.37
		pTVA	63.06 ± 6.64	54.17 ± 10.37	1.52	2.38	3.15E+00	38	0.00E+00	**	[3.17, 14.61]	9.90E-01	1.21E+01	0.87
		aTVA	55.00 ± 9.44	54.17 ± 10.37	2.17	2.38	2.60E-01	38	8.00E-01	ns	[-5.68, 7.35]	8.00E-02	3.17E-01	0.06
		TVAs	55.74 ± 9.88	54.17 ± 10.37	1.29	2.38	6.00E-01	78	5.50E-01	ns	[-3.64, 6.78]	1.60E-01	3.05E-01	0.09
Age	LIN	mTVA	53.85 ± 12.06	45.19 ± 13.11	2.41	2.62	2.43E+00	50	2.00E-02	*	[1.50, 15.81]	6.70E-01	2.95E+00	0.66
		pTVA	50.32 ± 10.46	45.19 ± 13.11	2.09	2.62	1.53E+00	50	1.30E-01	ns	[-1.61, 11.87]	4.20E-01	7.23E-01	0.32
		aTVA	55.77 ± 14.94	45.19 ± 13.11	2.99	2.62	2.66E+00	50	1.00E-02	*	[2.59, 18.56]	7.40E-01	4.65E+00	0.74
		TVAs	53.31 ± 12.82	45.19 ± 13.11	1.46	2.62	2.75E+00	102	1.00E-02	**	[2.27, 13.97]	6.20E-01	5.95E+00	0.78
	VLS	mTVA	52.88 ± 10.58	52.24 ± 9.68	2.12	1.94	2.20E-01	50	8.20E-01	ns	[-5.12, 6.40]	6.00E-02	2.84E-01	0.06
		pTVA	50.96 ± 11.40	52.24 ± 9.68	2.28	1.94	-4.30E-01	50	6.70E-01	ns	[-7.29, 4.73]	1.20E-01	3.00E-01	0.07
		aTVA	63.14 ± 11.82	52.24 ± 9.68	2.36	1.94	3.56E+00	50	0.00E+00	***	[4.76, 17.04]	9.90E-01	3.70E+01	0.94
		TVAs	55.66 ± 12.48	52.24 ± 9.68	1.42	1.94	1.26E+00	102	2.10E-01	ns	[-1.95, 8.79]	2.90E-01	4.67E-01	0.24
Identity	LIN	mTVA	49.70 ± 10.85	56.05 ± 9.41	2.63	2.28	-1.82E+00	34	8.00E-02	ns	[-13.43, 0.72]	6.10E-01	1.15E+00	0.43
		pTVA	54.38 ± 9.34	56.05 ± 9.41	2.27	2.28	-5.20E-01	34	6.10E-01	ns	[-8.21, 4.86]	1.70E-01	3.58E-01	0.08
		aTVA	40.50 ± 8.37	56.05 ± 9.41	2.03	2.28	-5.09E+00	34	0.00E+00	****	[-21.76, -9.35]	1.70E+00	1.25E+03	1.00
		TVAs	48.19 ± 11.18	56.05 ± 9.41	1.54	2.28	-2.65E+00	70	1.00E-02	*	[-13.79, -1.94]	7.20E-01	4.69E+00	0.74
	VLS	mTVA	46.29 ± 8.86	60.28 ± 14.19	2.15	3.44	-3.45E+00	34	0.00E+00	**	[-22.24, -5.75]	1.15E+00	2.21E+01	0.92
		pTVA	63.33 ± 6.75	60.28 ± 14.19	1.64	3.44	8.00E-01	34	4.30E-01	ns	[-4.69, 10.79]	2.70E-01	4.13E-01	0.12
		aTVA	51.01 ± 8.00	60.28 ± 14.19	1.94	3.44	-2.35E+00	34	2.00E-02	*	[-17.30, -1.24]	7.80E-01	2.53E+00	0.63
		TVAs	53.54 ± 10.69	60.28 ± 14.19	1.47	3.44	-2.09E+00	70	4.00E-02	*	[-13.16, -0.31]	5.70E-01	1.64E+00	0.54

Supplementary Table S21: Comparing the performance of the human listeners at discriminating speaker identity-related information by ROI. This table reports the significance of the A1-TVAs difference in the speaker identity categorization and discrimination performance. Two-sample t-tests were conducted between the scores of human listeners at discriminating the speaker gender (2 classes), age (2 classes), and identity (17 classes) of the 18 Test Stimuli that were reconstructed from the VLS features with those from LIN features. s.e.m. = standard error of the mean. Here are reported the results of the statistical tests, t-value, degree of freedom (dof), p-value, degree of significance (unc. sig.), 95% confidence interval (CI95%), effect size (Cohen-d), Bayes Factor (BF10), and statistical power (power) for each speaker identity information and ROI.

Supplementary Audio S1: Voice latent space interpolation.

The audio files are two original voice samples (A, B); the synthesized voice samples from the spectrograms of the autoencoder reconstructions of the original two voice samples (A', B'); the synthesized voice samples from the spectrograms of the linearly interpolated *voice latent space* (VLS; A_to_B; Figure 3.1c).

A.wav: Original voice sample of a female Chinese speaker
A'.wav: Voice sample A.wav reconstructed by the autoencoder
B.wav: Original voice sample of a male French speaker
B'.wav: Voice sample B.wav reconstructed by the autoencoder
A_to_B_l*x*.wav: Reconstructed voice samples from the linear interpolation between A and B VLS, where *x* is the interpolation step (0.2, 0.4, 0.6, 0.8).

Link:
https://drive.google.com/drive/folders/1WQonOiO_FpQvj9mT3okVSasno_rck_u3?usp=sharing

Supplementary Audio S2: Brain-based voice reconstructions.

The audio files are reconstructed voice samples from the fMRI responses in the speakers' temporal voice areas (TVAs). These sounds were used in the quantitative and subjective voice identity tests (Figure 3.4). The samples below are from a German and a Spanish speaker. The sounds are reconstructed for each speaker using 2 models: LIN and VLS.

example1_orig.wav: Original voice sample of a male German speaker
example1_VLS_RaTVA.wav: Reconstructed voice sample from fMRI activity in the right anterior temporal voice area (RaTVA) using the VLS model
example1_LIN_RaTVA.wav: Reconstructed voice sample from fMRI activity in the right anterior temporal voice area (RaTVA) using the LIN model

example2_orig.wav: Original voice sample of a male Spanish speaker
example2_VLS_LmTVA.wav: Reconstructed voice sample from fMRI activity in the

left middle voice area (LmTVA) using the VLS model

example2_LIN_LmTVA.wav: Reconstructed voice sample from fMRI activity in the left middle voice area (LmTVA) using the LIN model

Link:
https://drive.google.com/drive/folders/1AwAV2zigRb9DxDt_xhyea8sp13Zvxcuk?usp=share_link

Abbreviations

A1	primary auditory cortex
AF	arcuate fasciculus
ANN	artificial neural network
DNN	deep neural network
BOLD	blood-oxygen-level dependent
CV	conspecific vocalization
CNN	convolutional neural network
EC	effective connectivity
ECOG	electrocorticography
EEG	electroencephalography
ERP	event-related potential
FC	functional connectivity
FVA	frontal voice area
fMRI	functional magnetic resonance imaging
IFC	inferior frontal cortex
IFG	inferior frontal gyrus
MEG	magnetoencephalography
rTMS	repetitive transcranial magnetic stimulation
SEM	standard errors of the mean
SPM	statistical parametric mapping
sEEG	stereoelectroencephalography
STG	superior temporal gyrus
STS	superior temporal sulcus
TVA	temporal voice area
VA	voice area

References

Aboitiz, Francisco. 2018. « A Brain for Speech. Evolutionary Continuity in Primate and Human Auditory-Vocal Processing ». *Frontiers in Neuroscience* 12.

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, et Gael Varoquaux. 2014. « Machine learning for neuroimaging with scikit-learn ». *Frontiers in Neuroinformatics* 8.

Agamaite, James A., Chia-Jung Chang, Michael S. Osmanski, et Xiaoqin Wang. 2015. « A Quantitative Acoustic Analysis of the Vocal Repertoire of the Common Marmoset (*Callithrix Jacchus*) ». *The Journal of the Acoustical Society of America* 138(5):2906-28. doi: 10.1121/1.4934268.

Aglieri, Virginia, Bastien Cagna, Lionel Velly, Sylvain Takerkart, et Pascal Belin. 2021. « FMRI-Based Identity Classification Accuracy in Left Temporal and Frontal Regions Predicts Speaker Recognition Performance ». *Scientific Reports* 11(1):489. doi: 10.1038/s41598-020-79922-7.

Aglieri, Virginia, Thierry Chaminade, Sylvain Takerkart, et Pascal Belin. 2018. « Functional connectivity within the voice perception network and its behavioural relevance ». *Neuroimage* 183:356-65. doi: 10.1016/j.neuroimage.2018.08.011.

Agus, Trevor R., Sébastien Paquette, Clara Suied, Daniel Pressnitzer, et Pascal Belin. 2017. « Voice selectivity in the temporal voice area despite matched low-level acoustic cues ». *Scientific Reports* 7:11526. doi: 10.1038/s41598-017-11684-1.

Akbari, Hassan, Bahar Khalighinejad, Jose L. Herrero, Ashesh D. Mehta, et Nima Mesgarani. 2019. « Towards Reconstructing Intelligible Speech from the Human Auditory Cortex ». *Scientific Reports* 9(1):874. doi: 10.1038/s41598-018-37359-z.

Akuzawa, Kei, Yusuke Iwasawa, et Yutaka Matsuo. 2018. « Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder ». P. 3067-71 in *Interspeech 2018*. ISCA.

Albouy, Philippe, Lucas Benjamin, Benjamin Morillon, et Robert J. Zatorre. 2020. « Distinct Sensitivity to Spectrotemporal Modulation Supports Brain Asymmetry for Speech and Melody ». *Science* 367(6481):1043-47. doi: 10.1126/science.aaz3468.

Allen, Emily J., Philip C. Burton, Cheryl A. Olman, et Andrew J. Oxenham. 2017. « Representations of Pitch and Timbre Variation in Human Auditory Cortex ». *The Journal of Neuroscience* 37(5):1284-93. doi: 10.1523/JNEUROSCI.2336-16.2016.

Anderson, A. K., et E. A. Phelps. 2001. « Lesions of the Human Amygdala Impair Enhanced Perception of Emotionally Salient Events ». *Nature* 411(6835):305-9. doi: 10.1038/35077083.

Andics, Attila, James M. McQueen, Karl Magnus Petersson, Viktor Gál, Gábor Rudas, et Zoltán Vidnyánszky. 2010. « Neural Mechanisms for Voice Recognition ». *NeuroImage* 52(4):1528-40. doi: 10.1016/j.neuroimage.2010.05.048.

Anwander, A., M. Tittgemeyer, D. Y. von Cramon, A. D. Friederici, et T. R. Knösche. 2007. « Connectivity-Based Parcellation of Broca's Area ». *Cerebral Cortex (New York, N.Y.: 1991)* 17(4):816-25. doi: 10.1093/cercor/bhk034.

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, et Gregor Weber. 2020. « Common Voice: A Massively-Multilingual Speech Corpus ».

Arnal, Luc H., Adeen Flinker, Andreas Kleinschmidt, Anne-Lise Giraud, et David Poeppel. 2015. « Human Screams Occupy a Privileged Niche in the Communication Soundscape ». *Current Biology* 25(15):2051-56. doi: 10.1016/j.cub.2015.06.043.

Ashburner, John. 2012. « SPM: A History ». *NeuroImage* 62(2):791-800. doi: 10.1016/j.neuroimage.2011.10.025.

Averbeck, Bruno B., et Lizabeth M. Romanski. 2006. « Probabilistic Encoding of Vocalizations in Macaque Ventral Lateral Prefrontal Cortex ». *Journal of Neuroscience* 26(43):11023-33. doi: 10.1523/JNEUROSCI.3466-06.2006.

Balezeau, Fabien, Benjamin Wilson, Guillermo Gallardo, Fred Dick, William Hopkins, Alfred Anwander, Angela D. Friederici, Timothy D. Griffiths, et Christopher I. Petkov. 2020. « Primate Auditory Prototype in the Evolution of the Arcuate Fasciculus ». *Nature Neuroscience* 23(5):611-14. doi: 10.1038/s41593-020-0623-9.

Bando, Yoshiaki, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, et Tatsuya Kawahara. 2018. « Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization ». P. 716-20 in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Barbero, Francesca M., Roberta P. Calce, Siddharth Talwar, Bruno Rossion, et Olivier Collignon. 2021. « Fast Periodic Auditory Stimulation Reveals a Robust Categorical Response to Voices in the Human Brain ». *eNeuro* 8(3):ENEURO.0471-20.2021. doi: 10.1523/ENEURO.0471-20.2021.

Baumann, Simon, Christopher Petkov, et Timothy Griffiths. 2013. « A unified framework for the organization of the primate auditory cortex ». *Frontiers in Systems Neuroscience* 7.

Beecher, M. D., M. R. Petersen, S. R. Zoloth, D. B. Moody, et W. C. Stebbins. 1979. « Perception of Conspecific Vocalizations by Japanese Macaques. Evidence for Selective Attention and Neural Lateralization ». *Brain, Behavior and Evolution* 16(5-6):443-60. doi: 10.1159/000121881.

Belin, Pascal. 2017. « Similarities in Face and Voice Cerebral Processing ». *Visual Cognition* 25(4-6):658-65. doi: 10.1080/13506285.2017.1339156.

Belin, Pascal, Patricia E. G. Bestelmeyer, Marianne Latinus, et Rebecca Watson. 2011. « Understanding Voice Perception: Understanding Voice Perception ». *British Journal of Psychology* 102(4):711-25. doi: 10.1111/j.2044-8295.2011.02041.x.

Belin, Pascal, Clémentine Bodin, et Virginia Aglieri. 2018. « A "Voice Patch" System in the Primate Brain for Processing Vocal Information? » *Hearing Research* 366:65-74. doi: 10.1016/j.heares.2018.04.010.

Belin, Pascal, Shirley Fecteau, et Catherine Bédard. 2004. « Thinking the Voice: Neural Correlates of Voice Perception ». *Trends in Cognitive Sciences* 8(3):129-35. doi: 10.1016/j.tics.2004.01.008.

Belin, Pascal, et Robert J. Zatorre. 2003. « Adaptation to Speaker's Voice in Right Anterior Temporal Lobe: » *NeuroReport* 14(16):2105-9. doi: 10.1097/00001756-200311140-00019.

Belin, Pascal, Robert J. Zatorre, et Pierre Ahad. 2002. « Human Temporal-Lobe Response to Vocal Sounds ». *Cognitive Brain Research* 13(1):17-26. doi: 10.1016/S0926-6410(01)00084-2.

Belin, Pascal, Robert J. Zatorre, Philippe Lafaille, Pierre Ahad, et Bruce Pike. 2000. « Voice-Selective Areas in Human Auditory Cortex ». *Nature* 403(6767):309-12. doi: 10.1038/35002078.

Bell, Silvia M., et Mary D. Ainsworth. 1972. « Infant crying and maternal responsiveness ». *Child Development* 43(4):1171-90. doi: 10.2307/1127506.

Bendor, Daniel, et Xiaoqin Wang. 2008. « Neural Response Properties of Primary, Rostral, and Rostrotemporal Core Fields in the Auditory Cortex of Marmoset Monkeys ». *Journal of Neurophysiology* 100(2):888-906. doi: 10.1152/jn.00884.2007.

Bengio, Yoshua, Aaron Courville, et Pascal Vincent. 2014. « Representation Learning: A Review and New Perspectives ».

van den Berg, Jw., J. T. Zantema, et P. Doornenbal Jr. 1957. « On the Air Resistance and the Bernoulli Effect of the Human Larynx ». *The Journal of the Acoustical Society of America* 29(5):626-31. doi: 10.1121/1.1908987.

Besle, Julien, Olivier Mougin, Rosa-María Sánchez-Panchuelo, Cornelis Lanting, Penny Gowland, Richard Bowtell, Susan Francis, et Katrin Krumbholz. 2019. « Is Human Auditory Cortex Organization Compatible With the Monkey Model? Contrary Evidence From Ultra-High-Field Functional and Structural MRI ». *Cerebral Cortex (New York, N.Y.: 1991)* 29(1):410-28. doi: 10.1093/cercor/bhy267.

Best, Paul, Sébastien Paris, Hervé Glotin, et Ricard Marxer. 2023. « Deep Audio Embeddings for Vocalisation Clustering ». *PLOS ONE* 18(7):e0283396. doi: 10.1371/journal.pone.0283396.

Bestelmeyer, Patricia E. G., Pascal Belin, et Marie-Helene Grosbras. 2011. « Right temporal TMS impairs voice detection ». *Current Biology* 21(20):R838-39. doi: 10.1016/j.cub.2011.08.046.

Bestelmeyer, Patricia E. G., Marianne Latinus, Laetitia Bruckert, Julien Rouger, Frances Crabbe, et Pascal Belin. 2012. « Implicitly Perceived Vocal Attractiveness Modulates Prefrontal Cortex Activity ». *Cerebral Cortex* 22(6):1263-70. doi: 10.1093/cercor/bhr204.

Bestelmeyer, Patricia E. G., et Constanze Mühl. 2022. « Neural dissociation of the acoustic and cognitive representation of voice identity ». *NeuroImage* 263:119647. doi: 10.1016/j.neuroimage.2022.119647.

Bezerra, Bruna Martins, et Antonio Souto. 2008. « Structure and Usage of the Vocal Repertoire of Callithrix Jacchus ». *International Journal of Primatology* 29(3):671-701. doi: 10.1007/s10764-008-9250-0.

Bhattacharya, Gautam, Jahangir Alam, et Patrick Kenny. 2017. « Deep Speaker Embeddings for Short-Duration Speaker Verification ». P. 1517-21 in *Interspeech 2017*. ISCA.

Blaauw, Merlijn, et Jordi Bonada. 2016. « Modeling and Transforming Speech Using Variational Autoencoders ». P. 1770-74 in *Interspeech 2016*. ISCA.

Blank, Helen, Alfred Anwander, et Katharina Von Kriegstein. 2011. « Direct Structural Connections between Voice- and Face-Recognition Areas ». *The Journal of Neuroscience* 31(36):12906-15. doi: 10.1523/JNEUROSCI.2091-11.2011.

Blank, Helen, Nuri Wieland, et Katharina von Kriegstein. 2014. « Person Recognition and the Brain: Merging Evidence from Patients and Healthy Individuals ». *Neuroscience & Biobehavioral Reviews* 47:717-34. doi: 10.1016/j.neubiorev.2014.10.022.

Bodin, C., S. Takerkart, P. Belin, et O. Coulon. 2018. « Anatomo-Functional Correspondence in the Superior Temporal Sulcus ». *Brain Structure and Function* 223(1):221-32. doi: 10.1007/s00429-017-1483-2.

Bodin, Clémentine, et Pascal Belin. 2020. « Exploring the Cerebral Substrate of Voice Perception in Primate Brains ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 375(1789):20180386. doi: 10.1098/rstb.2018.0386.

Bodin, Clémentine, Régis Trapeau, Bruno Nazarian, Julien Sein, Xavier Degiovanni, Joël Baurberg, Emilie Rapha, Luc Renaud, Bruno L. Giordano, et Pascal Belin. 2021. « Functionally Homologous Representation of Vocalizations in the Auditory Cortex of Humans and Macaques ». *Current Biology* S0960982221011477. doi: 10.1016/j.cub.2021.08.043.

Boë, Louis-Jean, Frédéric Berthommier, Thierry Legou, Guillaume Captier, Caralyn Kemp, Thomas R. Sawallis, Yannick Becker, Arnaud Rey, et Joël Fagot. 2017. « Evidence of a Vocalic Proto-System in the Baboon (Papio Papio) Suggests Pre-Hominin Speech Precursors ». *PLOS ONE* 12(1):e0169321. doi: 10.1371/journal.pone.0169321.

Bonte, Milene, Lars Hausfeld, Wolfgang Scharke, Giancarlo Valente, et Elia Formisano. 2014. « Task-Dependent Decoding of Speaker and Vowel Identity from Auditory Cortical Response Patterns ».

The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 34(13):4548-57. doi: 10.1523/JNEUROSCI.4339-13.2014.

Bouthillier, Xavier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, et Pascal Vincent. 2021. « Accounting for Variance in Machine Learning Benchmarks ».

Brewer, Alyssa A., et Brian Barton. 2016. « Maps of the Auditory Cortex ». *Annual Review of Neuroscience* 39(1):385-407. doi: 10.1146/annurev-neuro-070815-014045.

Brookes, Matthew J., James Leggett, Molly Rea, Ryan M. Hill, Niall Holmes, Elena Boto, et Richard Bowtell. 2022. « Magnetoencephalography with Optically Pumped Magnetometers (OPM-MEG): The next Generation of Functional Neuroimaging ». *Trends in Neurosciences* 45(8):621-34. doi: 10.1016/j.tins.2022.05.008.

Bryant, Katherine, et Todd Preuss. 2018. « A Comparative Perspective on the Human Temporal Lobe ». P. 239-58 in.

Buckner, Randy L., et Fenna M. Krienen. 2013. « The Evolution of Distributed Association Networks in the Human Brain ». *Trends in Cognitive Sciences* 17(12):648-65. doi: 10.1016/j.tics.2013.09.017.

Cadena, Santiago A., George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, et Alexander S. Ecker. 2019. « Deep Convolutional Models Improve Predictions of Macaque V1 Responses to Natural Images ». *PLOS Computational Biology* 15(4):e1006897. doi: 10.1371/journal.pcbi.1006897.

Cammoun, Leila, Jean Philippe Thiran, Alessandra Griffa, Reto Meuli, Patric Hagmann, et Stephanie Clarke. 2015. « Intrahemispheric Cortico-Cortical Connections of the Human Auditory Cortex ». *Brain Structure and Function* 220(6):3537-53. doi: 10.1007/s00429-014-0872-z.

Campanella, Salvatore, et Pascal Belin. 2007. « Integrating Face and Voice in Person Perception ». *Trends in Cognitive Sciences* 11(12):535-43. doi: 10.1016/j.tics.2007.10.001.

Capilla, A., P. Belin, et J. Gross. 2013. « The Early Spatio-Temporal Correlates and Task Independence of Cerebral Voice Processing Studied with MEG ». *Cerebral Cortex* 23(6):1388-95. doi: 10.1093/cercor/bhs119.

Caucheteux, Charlotte, Alexandre Gramfort, et Jean-Rémi King. 2022. « Deep Language Algorithms Predict Semantic Comprehension from Brain Activity ». *Scientific Reports* 12(1):16327. doi: 10.1038/s41598-022-20460-9.

Caucheteux, Charlotte, Alexandre Gramfort, et Jean-Rémi King. 2023. « Evidence of a Predictive Coding Hierarchy in the Human Brain Listening to Speech ». *Nature Human Behaviour* 1-12. doi: 10.1038/s41562-022-01516-2.

Caucheteux, Charlotte, et Jean-Rémi King. 2022. « Brains and Algorithms Partially Converge in Natural Language Processing ». *Communications Biology* 5(1):1-10. doi: 10.1038/s42003-022-03036-1.

Ceravolo, Leonardo, Sascha Frühholz, Jordan Pierce, Didier Grandjean, et Julie Péron. 2021. « Basal Ganglia and Cerebellum Contributions to Vocal Emotion Processing as Revealed by High-Resolution fMRI ». *Scientific Reports* 11(1):10645. doi: 10.1038/s41598-021-90222-6.

Chang, Le, et Doris Y. Tsao. 2017. « The Code for Facial Identity in the Primate Brain ». *Cell* 169(6):1013-1028.e14. doi: 10.1016/j.cell.2017.05.011.

Chang, Nadine, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, et Elissa M. Aminoff. 2019. « BOLD5000, a Public fMRI Dataset While Viewing 5000 Visual Images ». *Scientific Data* 6(1):49. doi: 10.1038/s41597-019-0052-3.

Charest, I., C. Pernet, M. Latinus, F. Crabbe, et P. Belin. 2013. « Cerebral Processing of Voice Gender Studied Using a Continuous Carryover fMRI Design ». *Cerebral Cortex* 23(4):958-66. doi: 10.1093/cercor/bhs090.

Charest, Ian, Cyril R. Pernet, Guillaume A. Rousselet, Ileana Quiñones, Marianne Latinus, Sarah Fillion-Bilodeau, Jean-Pierre Chartrand, et Pascal Belin. 2009. « Electrophysiological evidence for an early processing of human voices ». *BMC Neuroscience* 10(1):127. doi: 10.1186/1471-2202-10-127.

Chen, Zijiao, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, et Juan Helen Zhou. 2023. « Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding ». P. 22710-20 in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE.

Choi, Jung Yoon, Daniel Y. Takahashi, et Asif A. Ghazanfar. 2015. « Cooperative vocal control in marmoset monkeys via vocal feedback ». *Journal of Neurophysiology* 114(1):274-83. doi: 10.1152/jn.00228.2015.

Christison-Lagay, Kate L., Sharath Bennur, Jennifer Blackwell, Jung H. Lee, Tim Schroeder, et Yale E. Cohen. 2014. « Natural variability in species-specific vocalizations constrains behavior and neural activity ». *Hearing research* 312:128-42. doi: 10.1016/j.heares.2014.03.007.

Cohen, Y. E., B. E. Russ, S. J. Davis, A. E. Baker, A. L. Ackelson, et R. Nitecki. 2009. « A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition ». *Proceedings of the National Academy of Sciences* 106(47):20045-50. doi: 10.1073/pnas.0907248106.

Collobert, Ronan, et Jason Weston. 2008. *A unified architecture for natural language processing: Deep neural networks with multitask learning*.

Conant, David, Kristofer E. Bouchard, et Edward F. Chang. 2014. « Speech Map in the Human Ventral Sensory-Motor Cortex ». *Current opinion in neurobiology* 0:63-67. doi: 10.1016/j.conb.2013.08.015.

Cope, Thomas E., Yury Shtyrov, Lucy J. MacGregor, Rachel Holland, Friedemann Pulvermüller, James B. Rowe, et Karalyn Patterson. 2020. « Anterior temporal lobe is necessary for efficient lateralised processing of spoken word identity ». *Cortex; a journal devoted to the study of the nervous system and behavior* 126:107-18. doi: 10.1016/j.cortex.2019.12.025.

Cordeau, Mélina, Ihsane Bichoutar, David Meunier, Kep-Kee Loh, Isaure Michaud, Olivier Coulon, Guillaume Auzias, et Pascal Belin. 2023. « Anatomo-functional correspondence in the voice-selective regions of human prefrontal cortex ». *NeuroImage* 279:120336. doi: 10.1016/j.neuroimage.2023.120336.

Cowen, Alan S., Marvin M. Chun, et Brice A. Kuhl. 2014. « Neural Portraits of Perception: Reconstructing Face Images from Evoked Brain Activity ». *NeuroImage* 94:12-22. doi: 10.1016/j.neuroimage.2014.03.018.

Cox, Robert W., et James S. Hyde. 1997. « Software Tools for Analysis and Visualization of fMRI Data ». *NMR in Biomedicine* 10(4-5):171-78. doi: 10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L.

Da Costa, S., W. van der Zwaag, J. P. Marques, R. S. J. Frackowiak, S. Clarke, et M. Saenz. 2011. « Human Primary Auditory Cortex Follows the Shape of Heschl's Gyrus ». *Journal of Neuroscience* 31(40):14067-75. doi: 10.1523/JNEUROSCI.2000-11.2011.

Dado, Thirza, Yağmur Güçlütürk, Luca Ambrogioni, Gabriëlle Ras, Sander Bosch, Marcel van Gerven, et Umut Güçlü. 2022. « Hyperrealistic Neural Decoding for Reconstructing Faces from fMRI Activations via the GAN Latent Space ». *Scientific Reports* 12(1):141. doi: 10.1038/s41598-021-03938-w.

Défossez, Alexandre, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, et Jean-Rémi King. 2022. « Decoding speech from non-invasive brain recordings ».

Diedrichsen, Jörn, et Nikolaus Kriegeskorte. 2017. « Representational Models: A Common Framework for Understanding Encoding, Pattern-Component, and Representational-Similarity Analysis ». *PLOS Computational Biology* 13(4):e1005508. doi: 10.1371/journal.pcbi.1005508.

Doersch, Carl. 2021. « Tutorial on Variational Autoencoders ».
Domínguez-Borràs, Judith, Raphaël Guex, Constantino Méndez-Bértolo, Guillaume Legendre, Laurent Spinelli, Stephan Moratti, Sascha Frühholz, Pierre Mégevand, Luc Arnal, Bryan Strange, Margitta Seeck, et Patrik Vuilleumier. 2019. « Human amygdala response to unisensory and multisensory emotion input: No evidence for superadditivity from intracranial recordings ». *Neuropsychologia* 131:9-24. doi: 10.1016/j.neuropsychologia.2019.05.027.
Dosenbach, Nico U. F., Jonathan M. Koller, Eric A. Earl, Oscar Miranda-Dominguez, Rachel L. Klein, Andrew N. Van, Abraham Z. Snyder, Bonnie J. Nagel, Joel T. Nigg, Annie L. Nguyen, Victoria Wesevich, Deanna J. Greene, et Damien A. Fair. 2017. « Real-Time Motion Analytics during Brain MRI Improve Data Quality and Reduce Costs ». *NeuroImage* 161:80-93. doi: 10.1016/j.neuroimage.2017.08.025.
Eichert, Nicole, Lennart Verhagen, Davide Folloni, Saad Jbabdi, Alexandre A. Khrapitchev, Nicola R. Sibson, Dante Mantini, Jerome Sallet, et Rogier B. Mars. 2019. « What Is Special about the Human Arcuate Fasciculus? Lateralization, Projections, and Expansion ». *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior* 118:107-15. doi: 10.1016/j.cortex.2018.05.005.
Eliades, Steven J., et Joji Tsunada. 2019. « Marmosets in Auditory Research ». P. 451-75 in *The Common Marmoset in Captivity and Biomedical Research*. Elsevier.
Epple, Gisela. 1968. « Comparative Studies on Vocalization in Marmoset Monkeys (Hapalidae) ». *Folia Primatologica* 8(1):1-40. doi: 10.1159/000155129.
Erickson, Karen A. 2017. « Comprehensive Literacy Instruction, Interprofessional Collaborative Practice, and Students With Severe Disabilities ». *American Journal of Speech-Language Pathology* 26(2):193-205. doi: 10.1044/2017_AJSLP-15-0067.
Esling, Philippe, Axel Chemla–Romeu-Santos, et Adrien Bitton. 2018. « Generative Timbre Spaces with Variational Audio Synthesis ».
Everett, Daniel L. 2005. « Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language ». *Current Anthropology* 46(4):621-46. doi: 10.1086/431525.
Fant, G. 1960. « Acoustic Theory Of Speech Production ».
Fecteau, Shirley, Jorge L. Armony, Yves Joanette, et Pascal Belin. 2004. « Is Voice Processing Species-Specific in Human Auditory Cortex? An fMRI Study ». *NeuroImage* 23(3):840-48. doi: 10.1016/j.neuroimage.2004.09.019.
Fecteau, Shirley, Jorge L. Armony, Yves Joanette, et Pascal Belin. 2005. « Sensitivity to Voice in Human Prefrontal Cortex ». *Journal of Neurophysiology* 94(3):2251-54. doi: 10.1152/jn.00329.2005.
Federer, Callie, Haoyan Xu, Alona Fyshe, et Joel Zylberberg. 2020. « Improved Object Recognition Using Neural Networks Trained to Mimic the Brain's Statistical Properties ». *Neural Networks* 131:103-14. doi: 10.1016/j.neunet.2020.07.013.
Fischer, Julia. 2021. « Primate Vocal Communication and the Evolution of Speech ». *Current Directions in Psychological Science* 30(1):55-60. doi: 10.1177/0963721420979580.
Fitch, W. T. 1997. « Vocal Tract Length and Formant Frequency Dispersion Correlate with Body Size in Rhesus Macaques ». *The Journal of the Acoustical Society of America* 102(2 Pt 1):1213-22. doi: 10.1121/1.421048.
Fitch, W. Tecumseh. 2000. « The Evolution of Speech: A Comparative Review ». *Trends in Cognitive Sciences* 4(7):258-67. doi: 10.1016/S1364-6613(00)01494-7.
Fitch, W. Tecumseh, Bart de Boer, Neil Mathur, et Asif A. Ghazanfar. 2016. « Monkey vocal tracts are speech-ready ». *Science Advances* 2(12):e1600723. doi: 10.1126/sciadv.1600723.

Fitch, W. Tecumseh, et Jonathan B. Fritz. 2006. « Rhesus Macaques Spontaneously Perceive Formants in Conspecific Vocalizations ». *The Journal of the Acoustical Society of America* 120(4):2132-41. doi: 10.1121/1.2258499.

Fitch, W. Tecumseh, et Marc D. Hauser. 1995. « Vocal Production in Nonhuman Primates: Acoustics, Physiology, and Functional Constraints on "Honest" Advertisement ». *American Journal of Primatology* 37(3):191-219. doi: 10.1002/ajp.1350370303.

Flinker, Adeen, Werner K. Doyle, Ashesh D. Mehta, Orrin Devinsky, et David Poeppel. 2019. « Spectrotemporal Modulation Provides a Unifying Framework for Auditory Cortical Asymmetries ». *Nature Human Behaviour* 3(4):393-405. doi: 10.1038/s41562-019-0548-z.

Formisano, Elia, Dae Shik Kim, Francesco Di Salle, Pierre Francois van de Moortele, Kamil Ugurbil, et Rainer Goebel. 2003. « Mirror-Symmetric Tonotopic Maps in Human Primary Auditory Cortex ». *Neuron* 40(4):859-69. doi: 10.1016/s0896-6273(03)00669-x.

Freiwald, Winrich A., Doris Y. Tsao, et Margaret S. Livingstone. 2009. « A Face Feature Space in the Macaque Temporal Lobe ». *Nature Neuroscience* 12(9):1187-96. doi: 10.1038/nn.2363.

Frey, Stephen, Jennifer S. W. Campbell, G. Bruce Pike, et Michael Petrides. 2008. « Dissociating the Human Language Pathways with High Angular Resolution Diffusion Fiber Tractography ». *Journal of Neuroscience* 28(45):11435-44. doi: 10.1523/JNEUROSCI.2388-08.2008.

Frey, Stephen, Scott Mackey, et Michael Petrides. 2014. « Cortico-Cortical Connections of Areas 44 and 45B in the Macaque Monkey ». *Brain and Language* 131:36-55. doi: 10.1016/j.bandl.2013.05.005.

Friston, K. J., D. E. Glaser, R. N. A. Henson, S. Kiebel, C. Phillips, et J. Ashburner. 2002. « Classical and Bayesian Inference in Neuroimaging: Applications ». *NeuroImage* 16(2):484-512. doi: 10.1006/nimg.2002.1091.

Friston, K. J., A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, et R. S. J. Frackowiak. 1994. « Statistical Parametric Maps in Functional Imaging: A General Linear Approach ». *Human Brain Mapping* 2(4):189-210. doi: 10.1002/hbm.460020402.

Frühholz, Sascha, et Pascal Belin. 2018. *The Oxford Handbook of Voice Perception*. OUP Oxford.

Frühholz, Sascha, Leonardo Ceravolo, et Didier Grandjean. 2012. « Specific Brain Networks during Explicit and Implicit Decoding of Emotional Prosody ». *Cerebral Cortex (New York, N.Y.: 1991)* 22(5):1107-17. doi: 10.1093/cercor/bhr184.

Frühholz, Sascha, et Didier Grandjean. 2013. « Amygdala subregions differentially respond and rapidly adapt to threatening voices ». *Cortex* 49(5):1394-1403. doi: 10.1016/j.cortex.2012.08.003.

Frühholz, Sascha, Christoph Hofstetter, Chiara Cristinzio, Arnaud Saj, Margitta Seeck, Patrik Vuilleumier, et Didier Grandjean. 2015. « Asymmetrical Effects of Unilateral Right or Left Amygdala Damage on Auditory Cortical Processing of Vocal Emotions ». *Proceedings of the National Academy of Sciences* 112(5):1583-88. doi: 10.1073/pnas.1411315112.

Fugate, Jennifer, Harold Gouzoules, et Lynne Nygaard. 2008. « Recognition of rhesus macaque (Macaca mulatta) noisy screams: Evidence from conspecifics and human listeners ». *American journal of primatology* 70:594-604. doi: 10.1002/ajp.20533.

Fukushima, Makoto, Alex M. Doyle, Matthew P. Mullarkey, Mortimer Mishkin, et Bruno B. Averbeck. 2015. « Distributed Acoustic Cues for Caller Identity in Macaque Vocalization ». *Royal Society Open Science* 2(12). doi: 10.1098/rsos.150432.

Fukushima, Makoto, Richard C. Saunders, David A. Leopold, Mortimer Mishkin, et Bruno B. Averbeck. 2014. « Differential Coding of Conspecific Vocalizations in the Ventral Auditory Cortical Stream ». *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34(13):4665-76. doi: 10.1523/JNEUROSCI.3969-13.2014.

Fullerton, Barbara C., et Deepak N. Pandya. 2007. « Architectonic Analysis of the Auditory-Related Areas of the Superior Temporal Region in Human Brain ». *The Journal of Comparative Neurology* 504(5):470-98. doi: 10.1002/cne.21432.

Gaziv, Guy, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, et Michal Irani. 2022. « Self-Supervised Natural Image Reconstruction and Large-Scale Semantic Classification from Brain Activity ». *NeuroImage* 254:119121. doi: 10.1016/j.neuroimage.2022.119121.

Gemmeke, Jort F., Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, et Marvin Ritter. 2017. « Audio Set: An Ontology and Human-Labeled Dataset for Audio Events ». P. 776-80 in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE.

Ghazanfar, A. A., D. Smith-Rohrberg, et M. D. Hauser. 2001. « The Role of Temporal Cues in Rhesus Monkey Vocal Recognition: Orienting Asymmetries to Reversed Calls ». *Brain, Behavior and Evolution* 58(3):163-72. doi: 10.1159/000047270.

Ghazanfar, Asif A., et Steven J. Eliades. 2014. « The neurobiology of primate vocal communication ». *Current opinion in neurobiology* 0:128-35. doi: 10.1016/j.conb.2014.06.015.

Ghazanfar, Asif A., et Laurie R. Santos. 2004. « Primate Brains in the Wild: The Sensory Bases for Social Interactions ». *Nature Reviews. Neuroscience* 5(8):603-16. doi: 10.1038/nrn1473.

Ghazanfar, Asif A., Hjalmar K. Turesson, Joost X. Maier, Ralph van Dinther, Roy D. Patterson, et Nikos K. Logothetis. 2007. « Vocal-Tract Resonances as Indexical Cues in Rhesus Monkeys ». *Current Biology: CB* 17(5):425-30. doi: 10.1016/j.cub.2007.01.029.

Gil-da-Costa, Ricardo, Alex Martin, Marco A. Lopes, Monica Muñoz, Jonathan B. Fritz, et Allen R. Braun. 2006. « Species-Specific Calls Activate Homologs of Broca's and Wernicke's Areas in the Macaque ». *Nature Neuroscience* 9(8):1064-70. doi: 10.1038/nn1741.

Giordano, Bruno L., Michele Esposito, Giancarlo Valente, et Elia Formisano. 2023. « Intermediate Acoustic-to-Semantic Representations Link Behavioral and Neural Responses to Natural Sounds ». *Nature Neuroscience* 1-9. doi: 10.1038/s41593-023-01285-9.

Giordano, Bruno L., Caroline Whiting, Nikolaus Kriegeskorte, Sonja A. Kotz, Joachim Gross, et Pascal Belin. 2021. « The Representational Dynamics of Perceived Voice Emotions Evolve from Categories to Dimensions ». *Nature Human Behaviour*. doi: 10.1038/s41562-021-01073-0.

Glasser, Matthew F., Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, et David C. Van Essen. 2016. « A Multi-Modal Parcellation of Human Cerebral Cortex ». *Nature* 536(7615):171-78. doi: 10.1038/nature18933.

Glover, G. H., T. Q. Li, et D. Ress. 2000. « Image-Based Method for Retrospective Correction of Physiological Motion Effects in fMRI: RETROICOR ». *Magnetic Resonance in Medicine* 44(1):162-67. doi: 10.1002/1522-2594(200007)44:1<162::aid-mrm23>3.0.co;2-e.

Glover, Gary H. 1999. « Deconvolution of Impulse Response in Event-Related BOLD fMRI1 ». *NeuroImage* 9(4):416-29. doi: 10.1006/nimg.1998.0419.

Gouzoules, Sarah, Harold Gouzoules, et Peter Marler. 1984. « Rhesus monkey (Macaca mulatta) screams: Representational signalling in the recruitment of agonistic aid ». *Animal Behaviour* 32(1):182-93. doi: 10.1016/S0003-3472(84)80336-X.

Gregor, Karol, Ivo Danihelka, Alex Graves, Danilo Rezende, et Daan Wierstra. 2015. « DRAW: A Recurrent Neural Network For Image Generation ». P. 1462-71 in *Proceedings of the 32nd International Conference on Machine Learning*. PMLR.

Griffin, D. et Jae Lim. 1983. « Signal estimation from modified short-time Fourier transform ». P. 804-7 in *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 8. Boston, MASS, USA: Institute of Electrical and Electronics Engineers.

Grijseels, Dori M., Brendan J. Prendergast, Julia C. Gorman, et Cory T. Miller. 2023. « The Neurobiology of Vocal Communication in Marmosets ». *Annals of the New York Academy of Sciences* n/a(n/a). doi: 10.1111/nyas.15057.

Güçlü, Umut, Jordy Thielen, Michael Hanke, et Marcel van Gerven. 2016. « Brains on Beats ». in *Advances in Neural Information Processing Systems*. Vol. 29, édité par D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, et R. Garnett. Curran Associates, Inc.

Guenther, Frank. 2017. « Neuroimaging of the speech network ». *The Journal of the Acoustical Society of America* 141(5_Supplement):3559. doi: 10.1121/1.4987547.

Gultekin, Yasemin B., et Steffen R. Hage. 2017. « Limiting Parental Feedback Disrupts Vocal Development in Marmoset Monkeys ». *Nature Communications* 8(1):14046. doi: 10.1038/ncomms14046.

Gultekin, Yasemin B., et Steffen R. Hage. 2018. « Limiting Parental Interaction during Vocal Development Affects Acoustic Call Structure in Marmoset Monkeys ». *Science Advances* 4(4):eaar4012. doi: 10.1126/sciadv.aar4012.

Gultekin, Yasemin B., David G. C. Hildebrand, Kurt Hammerschmidt, et Steffen R. Hage. 2021. « High Plasticity in Marmoset Monkey Vocal Development from Infancy to Adulthood ». *Science Advances* 7(27):eabf2938. doi: 10.1126/sciadv.abf2938.

Gutierrez, Miren. 2021. « Algorithmic Gender Bias and Audiovisual Data: A Research Agenda ». *International Journal of Communication* 15:439-61.

Hackett, T. A., T. M. Preuss, et J. H. Kaas. 2001. « Architectonic Identification of the Core Region in Auditory Cortex of Macaques, Chimpanzees, and Humans ». *The Journal of Comparative Neurology* 441(3):197-222. doi: 10.1002/cne.1407.

Hackett, T. A., I. Stepniewska, et J. H. Kaas. 1998. « Thalamocortical Connections of the Parabelt Auditory Cortex in Macaque Monkeys ». *The Journal of Comparative Neurology* 400(2):271-86. doi: 10.1002/(sici)1096-9861(19981019)400:2<271::aid-cne8>3.0.co;2-6.

Hackett, Troy A. 2011. « Information Flow in the Auditory Cortical Network ». *Hearing Research* 271(1-2):133-46. doi: 10.1016/j.heares.2010.01.011.

Hackett, Troy A. 2015. « Anatomic Organization of the Auditory Cortex ». P. 27-53 in *Handbook of Clinical Neurology*. Vol. 129. Elsevier.

Hage, Steffen R., et Andreas Nieder. 2013. « Single Neurons in Monkey Prefrontal Cortex Encode Volitional Initiation of Vocalizations ». *Nature Communications* 4:2409. doi: 10.1038/ncomms3409.

Haigh, J. A., et J. S. Mason. 1993. « Robust Voice Activity Detection Using Cepstral Features ». P. 321-24 in *Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation*. Beijing, China: IEEE.

Hamilton, Liberty S. 2019. « The Asymmetric Auditory Cortex ». *Nature Human Behaviour* 3(4):327-28. doi: 10.1038/s41562-019-0582-x.

Hamzei, Farsin, Magnus-Sebastian Vry, Dorothee Saur, Volkmar Glauche, Markus Hoeren, Irina Mader, Cornelius Weiller, et Michel Rijntjes. 2016. « The Dual-Loop Model and the Human Mirror Neuron System: an Exploratory Combined fMRI and DTI Study of the Inferior Frontal Gyrus ». *Cerebral Cortex* 26(5):2215-24. doi: 10.1093/cercor/bhv066.

Han, Kuan, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, et Zhongming Liu. 2019. « Variational Autoencoder: An Unsupervised Model for Encoding and Decoding fMRI Activity in Visual Cortex ». *NeuroImage* 198:125-36. doi: 10.1016/j.neuroimage.2019.05.039.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández

del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, et Travis E. Oliphant. 2020. « Array Programming with NumPy ». *Nature* 585(7825):357-62. doi: 10.1038/s41586-020-2649-2.

Hauser, M. D., et K. Andersson. 1994. « Left Hemisphere Dominance for Processing Vocalizations in Adult, but Not Infant, Rhesus Monkeys: Field Experiments ». *Proceedings of the National Academy of Sciences of the United States of America* 91(9):3946-48. doi: 10.1073/pnas.91.9.3946.

Hauser, Marc D. 1991. « Sources of Acoustic Variation in Rhesus Macaque (Macaca Mulatta))Vocalizations ». *Ethology* 89(1):29-46. doi: 10.1111/j.1439-0310.1991.tb00291.x.

Hauser, Marc D., et Peter Marler. 1993. « Food-associated calls in rhesus macaques (Macaca mulatta): II. Costs and benefits of call production and suppression ». *Behavioral Ecology* 4(3):206-12. doi: 10.1093/beheco/4.3.206.

Haxby, J. V., E. A. Hoffman, et M. I. Gobbini. 2000. « The Distributed Human Neural System for Face Perception ». *Trends in Cognitive Sciences* 4(6):223-33. doi: 10.1016/s1364-6613(00)01482-0.

Hershey, Shawn, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, et Kevin Wilson. 2017. « CNN architectures for large-scale audio classification ». P. 131-35 in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hesse, Janis K., et Doris Y. Tsao. 2020. « The Macaque Face Patch System: A Turtle's Underbelly for the Brain ». *Nature Reviews Neuroscience* 21(12):695-716. doi: 10.1038/s41583-020-00393-w.

Hickok, Gregory, et David Poeppel. 2004. « Dorsal and Ventral Streams: A Framework for Understanding Aspects of the Functional Anatomy of Language ». *Cognition* 92(1-2):67-99. doi: 10.1016/j.cognition.2003.10.011.

Hickok, Gregory, et David Poeppel. 2007. « The Cortical Organization of Speech Processing ». *Nature Reviews Neuroscience* 8(5):393-402. doi: 10.1038/nrn2113.

Higgins, Irina, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, et Matthew Botvinick. 2021. « Unsupervised Deep Learning Identifies Semantic Disentanglement in Single Inferotemporal Face Patch Neurons ». *Nature Communications* 12(1):6456. doi: 10.1038/s41467-021-26751-5.

Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, et Alexander Lerchner. 2017. « β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK ».

Hinton, G. E., et R. R. Salakhutdinov. 2006. « Reducing the Dimensionality of Data with Neural Networks ». *Science* 313(5786):504-7. doi: 10.1126/science.1127647.

Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, et Brian Kingsbury. 2012. « Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups ». *IEEE Signal Processing Magazine* 29(6):82-97. doi: 10.1109/MSP.2012.2205597.

Holland, Dominic, Joshua M. Kuperman, et Anders M. Dale. 2010. « Efficient correction of inhomogeneous static magnetic field-induced distortion in Echo Planar Imaging ». *NeuroImage* 50(1):175-83. doi: 10.1016/j.neuroimage.2009.11.044.

Horikawa, Tomoyasu, et Yukiyasu Kamitani. 2017. « Generic Decoding of Seen and Imagined Objects Using Hierarchical Visual Features ». *Nature Communications* 8(1):15037. doi: 10.1038/ncomms15037.

Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, et Hartwig Adam. 2017. « MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications ». *arXiv:1704.04861 [Cs]*.

Hsu, Wei-Ning, Yu Zhang, et James Glass. 2017. « Learning Latent Representations for Speech Generation and Transformation ». P. 1273-77 in *Interspeech 2017*. ISCA.

Isnard, Vincent. 2016. « L'efficacité du système auditif humain pour la reconnaissance de sons naturels ». These de doctorat, Paris 6.

Jafari, Azadeh, Audrey Dureux, Alessandro Zanini, Ravi S. Menon, Kyle M. Gilbert, et Stefan Everling. 2023. « A Vocalization-Processing Network in Marmosets ». *Cell Reports* 42(5):112526. doi: 10.1016/j.celrep.2023.112526.

Joly, Olivier, Christophe Pallier, Franck Ramus, Daniel Pressnitzer, Wim Vanduffel, et Guy A. Orban. 2012. « Processing of Vocalizations in Humans and Monkeys: A Comparative fMRI Study ». *NeuroImage* 62(3):1376-89. doi: 10.1016/j.neuroimage.2012.05.070.

Jovanovic, Vladimir, Adam Ryan Fishbein, Lisa de la Mothe, Kuo-Fen Lee, et Cory Thomas Miller. 2022. « Behavioral Context Affects Social Signal Representations within Single Primate Prefrontal Cortex Neurons ». *Neuron* 110(8):1318-1326.e4. doi: 10.1016/j.neuron.2022.01.020.

Kaas, Jon H., et Troy A. Hackett. 2000. « Subdivisions of auditory cortex and processing streams in primates ». *Proceedings of the National Academy of Sciences* 97(22):11793-99. doi: 10.1073/pnas.97.22.11793.

Kaas, Jon H., Troy A. Hackett, et Mark Jude Tramo. 1999. « Auditory processing in primate cerebral cortex ». *Current Opinion in Neurobiology* 9(2):164-70. doi: 10.1016/S0959-4388(99)80022-1.

Kajikawa, Yoshinao, Lisa A. de la Mothe, Suzanne Blumell, Susanne J. Sterbing-D'Angelo, William D'Angelo, Corrie R. Camalier, et Troy A. Hackett. 2008. « Coding of FM Sweep Trains and Twitter Calls in Area CM of Marmoset Auditory Cortex ». *Hearing Research* 239(1-2):107-25. doi: 10.1016/j.heares.2008.01.015.

Kanwisher, Nancy, Josh McDermott, et Marvin M. Chun. 1997. « The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception ». *The Journal of Neuroscience* 17(11):4302-11. doi: 10.1523/JNEUROSCI.17-11-04302.1997.

Kasper, Lars, Steffen Bollmann, Andreea O. Diaconescu, Chloe Hutton, Jakob Heinzle, Sandra Iglesias, Tobias U. Hauser, Miriam Sebold, Zina-Mary Manjaly, Klaas P. Pruessmann, et Klaas E. Stephan. 2017. « The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data ». *Journal of Neuroscience Methods* 276:56-72. doi: 10.1016/j.jneumeth.2016.10.019.

Kell, Alexander J. E., Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, et Josh H. McDermott. 2018. « A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy ». *Neuron* 98(3):630-644.e16. doi: 10.1016/j.neuron.2018.03.044.

Khaligh-Razavi, Seyed-Mahdi, et Nikolaus Kriegeskorte. 2014. « Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation » édité par J. Diedrichsen. *PLoS Computational Biology* 10(11):e1003915. doi: 10.1371/journal.pcbi.1003915.

Kikuchi, Yukiko, Barry Horwitz, et Mortimer Mishkin. 2010. « Hierarchical Auditory Processing Directed Rostrally along the Monkey's Supratemporal Plane ». *Journal of Neuroscience* 30(39):13021-30. doi: 10.1523/JNEUROSCI.2267-10.2010.

Kim, Boklye, Jennifer L. Boes, Peyton H. Bland, Thomas L. Chenevert, et Charles R. Meyer. 1999. « Motion Correction in fMRI via Registration of Individual Slices into an Anatomical Volume ». *Magnetic Resonance in Medicine* 41(5):964-72. doi: 10.1002/(SICI)1522-2594(199905)41:5<964::AID-MRM16>3.0.CO;2-D.

Kingma, Diederik P., et Jimmy Ba. 2015. « Adam: A Method for Stochastic Optimization ».

Kingma, Diederik P., et Max Welling. 2014. « Auto-Encoding Variational Bayes ». *arXiv:1312.6114 [Cs, Stat]*.

Kingma, Diederik P., et Max Welling. 2019. « An Introduction to Variational Autoencoders ». *Foundations and Trends® in Machine Learning* 12(4):307-92. doi: 10.1561/2200000056.

Kriegeskorte, Nikolaus. 2008. « Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience ». *Frontiers in Systems Neuroscience*. doi: 10.3389/neuro.06.004.2008.

von Kriegstein, K., D. R. R. Smith, R. D. Patterson, S. J. Kiebel, et T. D. Griffiths. 2010. « How the Human Brain Recognizes Speech in the Context of Changing Speakers ». *Journal of Neuroscience* 30(2):629-38. doi: 10.1523/JNEUROSCI.2742-09.2010.

Kriegstein, Katharina V., et Anne-Lise Giraud. 2004a. « Distinct Functional Substrates along the Right Superior Temporal Sulcus for the Processing of Voices ». *NeuroImage* 22(2):948-55. doi: 10.1016/j.neuroimage.2004.02.020.

Kriegstein, Katharina V., et Anne-Lise Giraud. 2004b. « Distinct functional substrates along the right superior temporal sulcus for the processing of voices ». *NeuroImage* 22(2):948-55. doi: 10.1016/j.neuroimage.2004.02.020.

von Kriegstein, Katharina, Evelyn Eger, Andreas Kleinschmidt, et Anne Lise Giraud. 2003. « Modulation of Neural Responses to Speech by Directing Attention to Voices or Verbal Content ». *Cognitive Brain Research* 17(1):48-55. doi: 10.1016/S0926-6410(03)00079-X.

Krizhevsky, A. 2009. « Learning Multiple Layers of Features from Tiny Images ».

Krizhevsky, Alex, Ilya Sutskever, et Geoffrey E. Hinton. 2017. « ImageNet Classification with Deep Convolutional Neural Networks ». *Communications of the ACM* 60(6):84-90. doi: 10.1145/3065386.

Krubitzer, Leah, et Dianna M. Kahn. 2003. « Nature versus Nurture Revisited: An Old Idea with a New Twist ». *Progress in Neurobiology* 70(1):33-52. doi: 10.1016/s0301-0082(03)00088-1.

Kubilius, Jonas, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, et James J. DiCarlo. 2018. *CORnet: Modeling the Neural Mechanisms of Core Object Recognition*. *preprint*. Neuroscience. doi: 10.1101/408385.

Kulkarni, Tejas D., Will Whitney, Pushmeet Kohli, et Joshua B. Tenenbaum. 2015. « Deep Convolutional Inverse Graphics Network ».

Lane, Harlan, et Bernard Tranel. 1971. « The Lombard Sign and the Role of Hearing in Speech ». *Journal of Speech and Hearing Research* 14(4):677-709. doi: 10.1044/jshr.1404.677.

Latinus, Marianne, et Pascal Belin. 2011. « Human Voice Perception ». *Current Biology* 21(4):R143-45. doi: 10.1016/j.cub.2010.12.033.

Latinus, Marianne, Frances Crabbe, et Pascal Belin. 2011. « Learning-Induced Changes in the Cerebral Processing of Voice Identity ». *Cerebral Cortex (New York, N.Y.: 1991)* 21(12):2820-28. doi: 10.1093/cercor/bhr077.

Latinus, Marianne, Phil McAleer, Patricia E. G. Bestelmeyer, et Pascal Belin. 2013. « Norm-Based Coding of Voice Identity in Human Auditory Cortex ». *Current Biology* 23(12):1075-80. doi: 10.1016/j.cub.2013.04.055.

Lauer, Jessy, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie Weygandt Mathis, et Alexander Mathis. 2022. « Multi-Animal Pose Estimation, Identification and Tracking with DeepLabCut ». *Nature Methods* 19(4):496-504. doi: 10.1038/s41592-022-01443-0.

Lavan, Nadine. 2023. « The Time Course of Person Perception From Voices: A Behavioral Study ». *Psychological Science* 34(7):771-83. doi: 10.1177/09567976231161565.

Le, Lynn, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel van Gerven, et Umut Güçlü. 2022. « Brain2Pix: Fully convolutional naturalistic video frame reconstruction from brain activity ». *Frontiers in Neuroscience* 16.

LeCun, Yann, Yoshua Bengio, et Geoffrey Hinton. 2015. « Deep Learning ». *Nature* 521(7553):436-44. doi: 10.1038/nature14539.

Lee, Honglak, Roger Grosse, Rajesh Ranganath, et Andrew Y. Ng. 2009. « Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations ». P. 609-16 in *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal Quebec Canada: ACM.

Lee, Young-Eun, Seo-Hyun Lee, Sang-Ho Kim, et Seong-Whan Lee. 2023. « Towards Voice Reconstruction from EEG during Imagined Speech ».

Leglaive, Simon, Laurent Girin, et Radu Horaud. 2018. « A variance modeling framework based on variational autoencoders for speech enhancement ». P. 1-6 in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*.

Lieberman, Philip, Edmund S. Crelin, et Dennis H. Klatt. 1972. « Phonetic Ability and Related Anatomy of the Newborn and Adult Human, Neanderthal Man, and the Chimpanzee ». *American Anthropologist* 74(3):287-307.

Lieberman, Philip H., Dennis H. Klatt, et William H. Wilson. 1969. « Vocal Tract Limitations on the Vowel Repertoires of Rhesus Monkey and other Nonhuman Primates ». *Science* 164(3884):1185-87. doi: 10.1126/science.164.3884.1185.

Lillicrap, Timothy P., Adam Santoro, Luke Marris, Colin J. Akerman, et Geoffrey Hinton. 2020. « Backpropagation and the Brain ». *Nature Reviews Neuroscience* 21(6):335-46. doi: 10.1038/s41583-020-0277-3.

Liu, Dong, Yue Li, Jianping Lin, Houqiang Li, et Feng Wu. 2020. « Deep Learning-Based Video Coding: A Review and a Case Study ». *ACM Computing Surveys* 53(1):11:1-11:35. doi: 10.1145/3368405.

Loh, Kep Kee, Michael Petrides, William D. Hopkins, Emmanuel Procyk, et Céline Amiez. 2017. « Cognitive Control of Vocalizations in the Primate Ventrolateral-Dorsomedial Frontal (VLF-DMF) Brain Network ». *Neuroscience and Biobehavioral Reviews* 82:32-44. doi: 10.1016/j.neubiorev.2016.12.001.

Loh, Kep Kee, Emmanuel Procyk, Rémi Neveu, Franck Lamberton, William D. Hopkins, Michael Petrides, et Céline Amiez. 2020. « Cognitive Control of Orofacial Motor and Vocal Responses in the Ventrolateral and Dorsomedial Human Frontal Cortex ». *Proceedings of the National Academy of Sciences* 117(9):4994-5005. doi: 10.1073/pnas.1916459117.

Lowe, Matthew X., Yalda Mohsenzadeh, Benjamin Lahner, Ian Charest, Aude Oliva, et Santani Teng. 2021. « Cochlea to Categories: The Spatiotemporal Dynamics of Semantic Auditory Representations ». *Cognitive Neuropsychology* 38(7-8):468-89. doi: 10.1080/02643294.2022.2085085.

Luzzi, Simona, Michela Coccia, Gabriele Polonara, Carlo Reverberi, Gabriella Ceravolo, Mauro Silvestrini, Fabio Fringuelli, Sara Baldinelli, Leandro Provinciali, et Guido Gainotti. 2018. « Selective Associative Phonagnosia after Right Anterior Temporal Stroke ». *Neuropsychologia* 116(Pt B):154-61. doi: 10.1016/j.neuropsychologia.2017.05.016.

Maguinness, Corrina, Claudia Roswandowitz, et Katharina von Kriegstein. 2018. « Understanding the Mechanisms of Familiar Voice-Identity Recognition in the Human Brain ». *Neuropsychologia* 116:179-93. doi: 10.1016/j.neuropsychologia.2018.03.039.

Maris, Eric, et Robert Oostenveld. 2007. « Nonparametric Statistical Testing of EEG- and MEG-Data ». *Journal of Neuroscience Methods* 164(1):177-90. doi: 10.1016/j.jneumeth.2007.03.024.

Martin, F. N., et C. A. Champlin. 2000. « Reconsidering the Limits of Normal Hearing ». *Journal of the American Academy of Audiology* 11(2):64-66.

Mathias, Samuel R., et Katharina von Kriegstein. 2014. « How do we recognise who is speaking? » *Frontiers in Bioscience-Scholar* 6(1):92-109. doi: 10.2741/S417.

Mathis, Alexander, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, et Matthias Bethge. 2018. « DeepLabCut: Markerless Pose Estimation of User-Defined Body Parts with Deep Learning ». *Nature Neuroscience* 21(9):1281-89. doi: 10.1038/s41593-018-0209-y.

May, Brad, David B. Moody, et William C. Stebbins. 1988. « The significant features of Japanese macaque coo sounds: a psychophysical study ». *Animal Behaviour* 36(5):1432-44. doi: 10.1016/S0003-3472(88)80214-8.

McInnes, Leland, John Healy, et Steve Astels. 2017. « Hdbscan: Hierarchical Density Based Clustering ». *The Journal of Open Source Software* 2(11):205. doi: 10.21105/joss.00205.

McInnes, Leland, John Healy, Nathaniel Saul, et Lukas Großberger. 2018. « UMAP: Uniform Manifold Approximation and Projection ». *Journal of Open Source Software* 3(29):861. doi: 10.21105/joss.00861.

Mesgarani, Nima, Connie Cheung, Keith Johnson, et Edward F. Chang. 2014. « Phonetic Feature Encoding in Human Superior Temporal Gyrus ». *Science (New York, N.Y.)* 343(6174):1006-10. doi: 10.1126/science.1245994.

Mikl, M., et M. Gajdos. 2014. « 23. Statistical characteristics of event related and block design datasets ». *Clinical Neurophysiology* 125(5):e32. doi: 10.1016/j.clinph.2013.12.061.

Mikl, Michal, Radek Mareček, Petr Hluštík, Martina Pavlicová, Aleš Drastich, Pavel Chlebus, Milan Brázdil, et Petr Krupa. 2008. « Effects of spatial smoothing on fMRI group inferences ». *Magnetic Resonance Imaging* 26(4):490-503. doi: 10.1016/j.mri.2007.08.006.

Miller, Cory T., Kaylin Beck, Brooke Meade, et Xiaoqin Wang. 2009. « Antiphonal call timing in marmosets is behaviorally significant: Interactive playback experiments ». *Journal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiology* 195(8):783-89. doi: 10.1007/s00359-009-0456-1.

Miller, Cory T., Winrich A. Freiwald, David A. Leopold, Jude F. Mitchell, Afonso C. Silva, et Xiaoqin Wang. 2016. « Marmosets: A Neuroscientific Model of Human Social Behavior ». *Neuron* 90(2):219-33. doi: 10.1016/j.neuron.2016.03.018.

Miller, Cory T., Katherine Mandel, et Xiaoqin Wang. 2010. « The communicative content of the common marmoset phee call during antiphonal calling ». *American journal of primatology* 72(11):974-80. doi: 10.1002/ajp.20854.

Miller, Cory T., A. Wren Thomas, Samuel U. Nummela, et Lisa A. de la Mothe. 2015. « Responses of Primate Frontal Cortex Neurons during Natural Vocal Communication ». *Journal of Neurophysiology* 114(2):1158-71. doi: 10.1152/jn.01003.2014.

Miller, Paul, et Xiao-Jing Wang. 2006. « Inhibitory control by an integral feedback signal in prefrontal cortex: A model of discrimination between sequential stimuli ». *Proceedings of the National Academy of Sciences* 103(1):201-6. doi: 10.1073/pnas.0508072103.

Millet, Juliette, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, et Jean-Remi King. 2022. *Toward a realistic model of speech processing in the brain with self-supervised learning*. arXiv:2206.01685. arXiv. doi: 10.48550/arXiv.2206.01685.

Moerel, Michelle, Federico De Martino, et Elia Formisano. 2014. « An Anatomical and Functional Topography of Human Auditory Cortical Areas ». *Frontiers in Neuroscience* 8:225. doi: 10.3389/fnins.2014.00225.

Morel, A., P. E. Garraghty, et J. H. Kaas. 1993. « Tonotopic Organization, Architectonic Fields, and Connections of Auditory Cortex in Macaque Monkeys ». *The Journal of Comparative Neurology* 335(3):437-59. doi: 10.1002/cne.903350312.

Morillon, Benjamin, Luc H. Arnal, et Pascal Belin. 2022. « The Path of Voices in Our Brain ». *PLOS Biology* 20(7):e3001742. doi: 10.1371/journal.pbio.3001742.

Morosan, P., J. Rademacher, A. Schleicher, K. Amunts, T. Schormann, et K. Zilles. 2001. « Human Primary Auditory Cortex: Cytoarchitectonic Subdivisions and Mapping into a Spatial Reference System ». *NeuroImage* 13(4):684-701. doi: 10.1006/nimg.2000.0715.

de la Mothe, Lisa A., Suzanne Blumell, Yoshinao Kajikawa, et Troy A. Hackett. 2006. « Cortical Connections of the Auditory Cortex in Marmoset Monkeys: Core and Medial Belt Regions ». *The Journal of Comparative Neurology* 496(1):27-71. doi: 10.1002/cne.20923.

Mozafari, Milad, Leila Reddy, et Rufin VanRullen. 2020. « Reconstructing Natural Scenes from fMRI Patterns Using BigBiGAN ». *2020 International Joint Conference on Neural Networks (IJCNN)* 1-8. doi: 10.1109/IJCNN48605.2020.9206960.

Myers, Emily B., et Rachel M. Theodore. 2017. « Voice-Sensitive Brain Networks Encode Talker-Specific Phonetic Detail ». *Brain and Language* 165:33-44. doi: 10.1016/j.bandl.2016.11.001.

Naselaris, Thomas, Kendrick N. Kay, Shinji Nishimoto, et Jack L. Gallant. 2011. « Encoding and Decoding in fMRI ». *NeuroImage* 56(2):400-410. doi: 10.1016/j.neuroimage.2010.07.073.

Negus, V. E. 1950. « The Comparative Anatomy and Physiology of the Larynx ». *The Laryngoscope* 60(5):516-516. doi: 10.1288/00005537-195005000-00010.

Nielsen, Alan K. S., et Drew Rendall. 2018. « Comparative Perspectives on Communication in Human and Non-Human Primates: Grounding Meaning in Broadly Conserved Processes of Voice Production, Perception, Affect, and Cognition ». P. 0 in *The Oxford Handbook of Voice Perception*, édité par S. Frühholz et P. Belin. Oxford University Press.

Norcross, J. L., et John D. Newman. 1993. « Context and gender-specific differences in the acoustic structure of common marmoset (Callithrix jacchus) phee calls ». *American Journal of Primatology* 30(1):37-54. doi: 10.1002/ajp.1350300104.

Norman-Haignere, Sam V., Laura K. Long, Orrin Devinsky, Werner Doyle, Ifeoma Irobunda, Edward M. Merricks, Neil A. Feldstein, Guy M. McKhann, Catherine A. Schevon, Adeen Flinker, et Nima Mesgarani. 2022. « Multiscale Temporal Integration Organizes Hierarchical Computation in Human Auditory Cortex ». *Nature Human Behaviour* 1-15. doi: 10.1038/s41562-021-01261-y.

Okano, Hideyuki, Erika Sasaki, Tetsuo Yamamori, Atsushi Iriki, Tomomi Shimogori, Yoko Yamaguchi, Kiyoto Kasai, et Atsushi Miyawaki. 2016. « Brain/MINDS: A Japanese National Brain Project for Marmoset Neuroscience ». *Neuron* 92(3):582-90. doi: 10.1016/j.neuron.2016.10.018.

Oller, D. K., et R. E. Eilers. 1988. « The Role of Audition in Infant Babbling ». *Child Development* 59(2):441-49.

Orhan, Pierre, Yves Boubenec, et Jean-Rémi King. 2022. « Don't Stop the Training: Continuously-Updating Self-Supervised Algorithms Best Account for Auditory Responses in the Cortex ». *arXiv:2202.07290 [Cs, q-Bio]*.

Ortiz-Rios, Michael, Paweł Kuśmierek, Iain DeWitt, Denis Archakov, Frederico A. C. Azevedo, Mikko Sams, Iiro P. Jääskeläinen, Georgios A. Keliris, et Josef P. Rauschecker. 2015. « Functional MRI of the vocalization-processing network in the macaque brain ». *Frontiers in Neuroscience* 9:113. doi: 10.3389/fnins.2015.00113.

Osmanski, Michael S., et Xiaoqin Wang. 2011. « Measurement of Absolute Auditory Thresholds in the Common Marmoset (Callithrix Jacchus) ». *Hearing Research* 277(1-2):127-33. doi: 10.1016/j.heares.2011.02.001.

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

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang,

Junjie Bai, et Soumith Chintala. 2019. « PyTorch: An Imperative Style, High-Performance Deep Learning Library ».

Patterson, Karalyn, Peter J. Nestor, et Timothy T. Rogers. 2007. « Where Do You Know What You Know? The Representation of Semantic Knowledge in the Human Brain ». *Nature Reviews Neuroscience* 8(12):976-87. doi: 10.1038/nrn2277.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, et Édouard Duchesnay. 2018. « Scikit-learn: Machine Learning in Python ».

Pedregosa-Izquierdo, Fabian. 2015. « Feature Extraction and Supervised Learning on fMRI: From Practice to Theory ».

Penhune, V. B., R. J. Zatorre, J. D. MacDonald, et A. C. Evans. 1996. « Interhemispheric Anatomical Differences in Human Primary Auditory Cortex: Probabilistic Mapping and Volume Measurement from Magnetic Resonance Scans ». *Cerebral Cortex (New York, N.Y.: 1991)* 6(5):661-72. doi: 10.1093/cercor/6.5.661.

Pernet, Cyril R., Phil McAleer, Marianne Latinus, Krzysztof J. Gorgolewski, Ian Charest, Patricia E. G. Bestelmeyer, Rebecca H. Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, et Pascal Belin. 2015. « The Human Voice Areas: Spatial Organization and Inter-Individual Variability in Temporal and Extra-Temporal Cortices ». *NeuroImage* 119:164-74. doi: 10.1016/j.neuroimage.2015.06.050.

Perrodin, Catherine, Christoph Kayser, Taylor J. Abel, Nikos K. Logothetis, et Christopher I. Petkov. 2015. « Who Is That? Brain Networks and Mechanisms for Identifying Individuals ». *Trends in Cognitive Sciences* 19(12):783-96. doi: 10.1016/j.tics.2015.09.002.

Perrodin, Catherine, Christoph Kayser, Nikos K. Logothetis, et Christopher I. Petkov. 2011. « Voice cells in the primate temporal lobe ». *Current biology : CB* 21(16):1408-15. doi: 10.1016/j.cub.2011.07.028.

Petkov, Christopher I., Christoph Kayser, Thomas Steudel, Kevin Whittingstall, Mark Augath, et Nikos K. Logothetis. 2008. « A Voice Region in the Monkey Brain ». *Nature Neuroscience* 11(3):367-74. doi: 10.1038/nn2043.

Petkov, Christopher I., Nikos K. Logothetis, et Jonas Obleser. 2009. « Where Are the Human Speech and Voice Regions, and Do Other Animals Have Anything Like Them? » *The Neuroscientist* 15(5):419-29. doi: 10.1177/1073858408326430.

Petrides, Michael. 2014. *Neuroanatomy of Language Regions of the Human Brain*. First edition. Amsterdam: Elsevier/AP, Academic Press is an imprint of Elsevier.

Petrides, Michael, et Deepak N. Pandya. 2009. « Distinct Parietal and Temporal Pathways to the Homologues of Broca's Area in the Monkey ». *PLOS Biology* 7(8):e1000170. doi: 10.1371/journal.pbio.1000170.

Pisanski, Katarzyna, et Drew Rendall. 2011. « The Prioritization of Voice Fundamental Frequency or Formants in Listeners' Assessments of Speaker Size, Masculinity, and Attractiveness ». *The Journal of the Acoustical Society of America* 129(4):2201-12. doi: 10.1121/1.3552866.

Pistorio, Ashley L., Brett Vintch, et Xiaoqin Wang. 2006. « Acoustic Analysis of Vocal Development in a New World Primate, the Common Marmoset (Callithrix Jacchus)a) ». *J. Acoust. Soc. Am.* 120(3):16.

Plaut, Elad. 2018. « From Principal Subspaces to Principal Components with Linear Autoencoders ».

Poldrack, Russell A., Jeanette A. Mumford, et Thomas E. Nichols. 2011. *Handbook of Functional MRI Data Analysis*. Cambridge: Cambridge University Press.

Poremba, Amy, Megan Malloy, Richard C. Saunders, Richard E. Carson, Peter Herscovitch, et Mortimer Mishkin. 2004. « Species-Specific Calls Evoke Asymmetric Activity in the Monkey's Temporal Poles ». *Nature* 427(6973):448-51. doi: 10.1038/nature02268.

Prat, Yosef. 2019. « Animals Have No Language, and Humans Are Animals Too ». *Perspectives on Psychological Science* 14(5):885-93. doi: 10.1177/1745691619858402.

Prat, Yosef, Mor Taub, Ester Pratt, et Yossi Yovel. 2017. « An Annotated Dataset of Egyptian Fruit Bat Vocalizations across Varying Contexts and during Vocal Ontogeny ». *Scientific Data* 4(1):170143. doi: 10.1038/sdata.2017.143.

Rajan, Ramesh, Vladimir Dubaj, David H. Reser, et Marcello G. P. Rosa. 2013. « Auditory Cortex of the Marmoset Monkey - Complex Responses to Tones and Vocalizations under Opiate Anaesthesia in Core and Belt Areas ». *The European Journal of Neuroscience* 37(6):924-41. doi: 10.1111/ejn.12092.

Rauschecker, Josef. 2012. « Ventral and dorsal streams in the evolution of speech and language ». *Frontiers in Evolutionary Neuroscience* 4.

Rauschecker, Josef P. 1998. « Parallel Processing in the Auditory Cortex of Primates ». *Audiology and Neurotology* 3(2-3):86-103. doi: 10.1159/000013784.

Rauschecker, Josef P., et Sophie K. Scott. 2009. « Maps and Streams in the Auditory Cortex: Nonhuman Primates Illuminate Human Speech Processing ». *Nature Neuroscience* 12(6):718-24. doi: 10.1038/nn.2331.

Rauschecker, Josef P., et Biao Tian. 2000. « Mechanisms and streams for processing of "what" and "where" in auditory cortex ». *Proceedings of the National Academy of Sciences* 97(22):11800-806. doi: 10.1073/pnas.97.22.11800.

Remington, Evan D., Michael S. Osmanski, et Xiaoqin Wang. 2012. « An Operant Conditioning Method for Studying Auditory Behaviors in Marmoset Monkeys ». *PLOS ONE* 7(10):e47895. doi: 10.1371/journal.pone.0047895.

Rendall, D., M. J. Owren, et P. S. Rodman. 1998. « The Role of Vocal Tract Filtering in Identity Cueing in Rhesus Monkey (Macaca Mulatta) Vocalizations ». *The Journal of the Acoustical Society of America* 103(1):602-14. doi: 10.1121/1.421104.

Rendall, Drew, Peter S. Rodman, et Roger E. Emond. 1996. « Vocal recognition of individuals and kin in free-ranging rhesus monkeys ». *Animal Behaviour* 51(5):1007-15. doi: 10.1006/anbe.1996.0103.

Rendall, DREW, ROBERT M. Seyfarth, DOROTHY L. Cheney, et MICHAEL J. Owren. 1999. « The meaning and function of grunt variants in baboons ». *Animal Behaviour* 57(3):583-92. doi: 10.1006/anbe.1998.1031.

Rezende, Danilo Jimenez, Shakir Mohamed, et Daan Wierstra. 2014. « Stochastic Backpropagation and Approximate Inference in Deep Generative Models ». P. 1278-86 in *Proceedings of the 31st International Conference on Machine Learning*. PMLR.

Rilling, James, Matthew Glasser, Saad Jbabdi, Jesper Andersson, et Todd Preuss. 2012. « Continuity, Divergence, and the Evolution of Brain Language Pathways ». *Frontiers in Evolutionary Neuroscience* 3.

Rilling, James K., Matthew F. Glasser, Todd M. Preuss, Xiangyang Ma, Tiejun Zhao, Xiaoping Hu, et Timothy E. J. Behrens. 2008. « The Evolution of the Arcuate Fasciculus Revealed with Comparative DTI ». *Nature Neuroscience* 11(4):426-28. doi: 10.1038/nn2072.

Robinson, Bryan W. 1976. « Limbic Influences on Human Speech* ». *Annals of the New York Academy of Sciences* 280(1):761-71. doi: 10.1111/j.1749-6632.1976.tb25539.x.

Rocchi, Francesca, Hiroyuki Oya, Fabien Balezeau, Alexander J. Billig, Zsuzsanna Kocsis, Rick L. Jenison, Kirill V. Nourski, Christopher K. Kovach, Mitchell Steinschneider, Yukiko Kikuchi, Ariane E. Rhone, Brian J. Dlouhy, Hiroto Kawasaki, Ralph Adolphs, Jeremy D. W. Greenlee, Timothy D. Griffiths,

Matthew A. Howard, et Christopher I. Petkov. 2021. « Common Fronto-Temporal Effective Connectivity in Humans and Monkeys ». *Neuron*. doi: 10.1016/j.neuron.2020.12.026.

Roche, Fanny, Thomas Hueber, Maëva Garnier, Samuel Limier, et Laurent Girin. 2021. « Make That Sound More Metallic: Towards a Perceptually Relevant Control of the Timbre of Synthesizer Sounds Using a Variational Autoencoder ». *Transactions of the International Society for Music Information Retrieval* 4(1):52-66. doi: 10.5334/tismir.76.

Roche, Fanny, Thomas Hueber, Samuel Limier, et Laurent Girin. 2023. « Autoencoders for Music Sound Modeling: A Comparison of Linear, Shallow, Deep, Recurrent and Variational Models ». 8.

Romanski, Lizabeth M., et Bruno B. Averbeck. 2009. « The Primate Cortical Auditory System and Neural Representation of Conspecific Vocalizations ». *Annual review of neuroscience* 32:315-46. doi: 10.1146/annurev.neuro.051508.135431.

Romanski, Lizabeth M., Bruno B. Averbeck, et Mark Diltz. 2005. « Neural Representation of Vocalizations in the Primate Ventrolateral Prefrontal Cortex ». *Journal of Neurophysiology* 93(2):734-47. doi: 10.1152/jn.00675.2004.

Roswandowitz, Claudia, Huw Swanborough, et Sascha Frühholz. 2021. « Categorizing Human Vocal Signals Depends on an Integrated Auditory-Frontal Cortical Network ». *Human Brain Mapping* 42(5):1503-17. doi: 10.1002/hbm.25309.

Rowell, T. E., et R. A. Hinde. 1962. « Vocal Communication by the Rhesus Mojsxey (Macaca Mulatta) ». *Proceedings of the Zoological Society of London* 138(2):279-94. doi: 10.1111/j.1469-7998.1962.tb05698.x.

Rupp, Kyle, Jasmine L. Hect, Madison Remick, Avniel Ghuman, Bharath Chandrasekaran, Lori L. Holt, et Taylor J. Abel. 2022. « Neural Responses in Human Superior Temporal Cortex Support Coding of Voice Representations ». *PLOS Biology* 20(7):e3001675. doi: 10.1371/journal.pbio.3001675.

Ruthig, Philip, et Marc Schönwiesner. 2022. « Common Principles in the Lateralization of Auditory Cortex Structure and Function for Vocal Communication in Primates and Rodents ». *European Journal of Neuroscience* 55(3):827-45. doi: 10.1111/ejn.15590.

Rutz, Christian, Michael Bronstein, Aza Raskin, Sonja C. Vernes, Katherine Zacarian, et Damián E. Blasi. 2023. « Using Machine Learning to Decode Animal Communication ». *Science* 381(6654):152-55. doi: 10.1126/science.adg7314.

Rylands, Anthony B., éd. 1993. *Marmosets and Tamarins: Systematics, Behaviour, and Ecology*. Oxford, New York: Oxford University Press.

Sacks, HARVEY, EMANUEL A. Schegloff, et GAIL Jefferson. 1978. « chapter 1 - A Simplest Systematics for the Organization of Turn Taking for Conversation**This chapter is a variant version of "A Simplest Systematics for the Organization of Turn-Taking for Conversation," which was printed in Language, 50, 4 (1974), pp. 696–735. An earlier version of this paper was presented at the conference on "Sociology of Language and Theory of Speech Acts," held at the Centre for Interdisciplinary Research of the University of Bielefeld, Germany. We thank Dr. Anita Pomerantz and Mr. Richard Faumann for pointing out to us a number of errors in the text. » P. 7-55 in *Studies in the Organization of Conversational Interaction*, édité par J. Schenkein. Academic Press.

Sadagopan, Srivatsun, Nesibe Z. Temiz-Karayol, et Henning U. Voss. 2015. « High-Field Functional Magnetic Resonance Imaging of Vocalization Processing in Marmosets ». *Scientific Reports* 5:10950. doi: 10.1038/srep10950.

Sainburg, Tim, et Timothy Q. Gentner. 2021. « Toward a Computational Neuroethology of Vocal Communication: From Bioacoustics to Neurophysiology, Emerging Tools and Future Directions ». *Frontiers in Behavioral Neuroscience* 15:330. doi: 10.3389/fnbeh.2021.811737.

Sainburg, Tim, Brad Theilman, Marvin Thielk, et Timothy Q. Gentner. 2019. « Parallels in the Sequential Organization of Birdsong and Human Speech ». *Nature Communications* 10(1):3636. doi: 10.1038/s41467-019-11605-y.

Sainburg, Tim, Marvin Thielk, et Timothy Q. Gentner. 2020. « Finding, Visualizing, and Quantifying Latent Structure across Diverse Animal Vocal Repertoires » édité par F. E. Theunissen. *PLOS Computational Biology* 16(10):e1008228. doi: 10.1371/journal.pcbi.1008228.

Salimans, Tim, Diederik P. Kingma, et M. Welling. 2014. « Markov Chain Monte Carlo and Variational Inference: Bridging the Gap ».

Santoro, Roberta, Michelle Moerel, Federico De Martino, Giancarlo Valente, Kamil Ugurbil, Essa Yacoub, et Elia Formisano. 2017. « Reconstructing the Spectrotemporal Modulations of Real-Life Sounds from fMRI Response Patterns ». *Proceedings of the National Academy of Sciences* 114(18):4799-4804. doi: 10.1073/pnas.1617622114.

Schall, Sonja, Stefan J. Kiebel, Burkhard Maess, et Katharina Von Kriegstein. 2015. « Voice Identity Recognition: Functional Division of the Right STS and Its Behavioral Relevance ». *Journal of Cognitive Neuroscience* 27(2):280-91. doi: 10.1162/jocn_a_00707.

Schönwiesner, Marc, Peter Dechent, Dirk Voit, Christopher I. Petkov, et Katrin Krumbholz. 2015. « Parcellation of Human and Monkey Core Auditory Cortex with fMRI Pattern Classification and Objective Detection of Tonotopic Gradient Reversals ». *Cerebral Cortex (New York, N.Y.: 1991)* 25(10):3278-89. doi: 10.1093/cercor/bhu124.

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, et Evelina Fedorenko. 2021. « The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing ». *Proceedings of the National Academy of Sciences* 118(45):e2105646118. doi: 10.1073/pnas.2105646118.

Schrimpf, Martin, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, et James J. DiCarlo. 2018. *Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like? preprint*. Neuroscience. doi: 10.1101/407007.

Schuller, Björn. 2013. « Intelligent Audio Analysis ».

Schütt, Heiko H., Alexander D. Kipnis, Jörn Diedrichsen, et Nikolaus Kriegeskorte. 2021. « Statistical Inference on Representational Geometries ».

Schweinberger, Stefan R., Anja Herholz, et Werner Sommer. 1997. « Recognizing Famous Voices ». *Journal of Speech, Language, and Hearing Research* 40(2):453-63. doi: 10.1044/jslhr.4002.453.

Scott, Sophie K., Nadine Lavan, Sinead Chen, et Carolyn McGettigan. 2014. « The Social Life of Laughter ». *Trends in Cognitive Sciences* 18(12):618-20. doi: 10.1016/j.tics.2014.09.002.

Scotti, Paul S., Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, et Tanishq Mathew Abraham. 2023. « Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors ». doi: 10.48550/ARXIV.2305.18274.

Seyfarth, Robert M., et Dorothy L. Cheney. 2003. « Signalers and Receivers in Animal Communication ». *Annual Review of Psychology* 54:145-73. doi: 10.1146/annurev.psych.54.101601.145121.

Singh, Charu, Maarten Venter, Rajesh Kumar Muthu, et David Brown. 2018. « DSP-Based Voice Activity Detection and Background Noise Reduction ». *International Journal of Speech Technology* 21(4):851-59. doi: 10.1007/s10772-018-9556-z.

Sladky, Ronald, Karl J. Friston, Jasmin Tröstl, Ross Cunnington, Ewald Moser, et Christian Windischberger. 2011. « Slice-timing effects and their correction in functional MRI ». *NeuroImage* 58(2):588-94. doi: 10.1016/j.neuroimage.2011.06.078.

Smiley, John F., Troy A. Hackett, Todd M. Preuss, Cynthia Bleiwas, Khadija Figarsky, J. John Mann, Gorazd Rosoklija, Daniel C. Javitt, et Andrew J. Dwork. 2013. « Hemispheric asymmetry of primary auditory cortex and Heschl's gyrus in schizophrenia and nonpsychiatric brains ». *Psychiatry research* 214(3):10.1016/j.pscychresns.2013.08.009. doi: 10.1016/j.pscychresns.2013.08.009.

Smith, Eric Alden. 2010. « Communication and collective action: language and the evolution of human cooperation ». *Evolution and Human Behavior* 31(4):231-45. doi: 10.1016/j.evolhumbehav.2010.03.001.

Smith, Stephen M., Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E. J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, et Paul M. Matthews. 2004. « Advances in Functional and Structural MR Image Analysis and Implementation as FSL ». *NeuroImage* 23:S208-19. doi: 10.1016/j.neuroimage.2004.07.051.

Snoek, Lukas, Maite M. van der Miesen, Tinka Beemsterboer, Andries van der Leij, Annemarie Eigenhuis, et H. Steven Scholte. 2021. « The Amsterdam Open MRI Collection, a Set of Multimodal MRI Datasets for Individual Difference Analyses ». *Scientific Data* 8(1):85. doi: 10.1038/s41597-021-00870-6.

Sprung-Much, Trisanna, et Michael Petrides. 2018. « Morphological Patterns and Spatial Probability Maps of Two Defining Sulci of the Posterior Ventrolateral Frontal Cortex of the Human Brain: The Sulcus Diagonalis and the Anterior Ascending Ramus of the Lateral Fissure ». *Brain Structure & Function* 223(9):4125-52. doi: 10.1007/s00429-018-1733-y.

Sprung-Much, Trisanna, et Michael Petrides. 2020. « Morphology and Spatial Probability Maps of the Horizontal Ascending Ramus of the Lateral Fissure ». *Cerebral Cortex (New York, NY)* 30(3):1586-1602. doi: 10.1093/cercor/bhz189.

Staib, Matthias, et Sascha Frühholz. 2021. « Cortical Voice Processing Is Grounded in Elementary Sound Analyses for Vocalization Relevant Sound Patterns ». *Progress in Neurobiology* 200:101982. doi: 10.1016/j.pneurobio.2020.101982.

Staib, Matthias, et Sascha Frühholz. 2023. « Distinct functional levels of human voice processing in the auditory cortex ». *Cerebral Cortex* 33(4):1170-85. doi: 10.1093/cercor/bhac128.

Steiner, Florence, Marine Bobin, et Sascha Frühholz. 2021. « Auditory Cortical Micro-Networks Show Differential Connectivity during Voice and Speech Processing in Humans ». *Communications Biology* 4(1):1-10. doi: 10.1038/s42003-021-02328-2.

Steiner, Florence, Natalia Fernandez, Joris Dietziker, Philipp Stämpfli, Erich Seifritz, Anton Rey, et Sascha Frühholz. 2022. « Affective speech modulates a cortico-limbic network in real time ». *Progress in Neurobiology* 214:102278. doi: 10.1016/j.pneurobio.2022.102278.

Steiper, Michael E., et Erik R. Seiffert. 2012. « Evidence for a Convergent Slowdown in Primate Molecular Rates and Its Implications for the Timing of Early Primate Evolution ». *Proceedings of the National Academy of Sciences* 109(16):6006-11. doi: 10.1073/pnas.1119506109.

Stiebler, I., R. Neulist, I. Fichtel, et G. Ehret. 1997. « The Auditory Cortex of the House Mouse: Left-Right Differences, Tonotopic Organization and Quantitative Analysis of Frequency Representation ». *Journal of Comparative Physiology A: Sensory, Neural, and Behavioral Physiology* 181(6):559-71. doi: 10.1007/s003590050140.

Suied, Clara, Trevor R. Agus, Simon J. Thorpe, Nima Mesgarani, et Daniel Pressnitzer. 2014. « Auditory Gist: Recognition of Very Short Sounds from Timbre Cues ». *The Journal of the Acoustical Society of America* 135(3):1380-91. doi: 10.1121/1.4863659.

Sweet, Robert A., Karl-Anton Dorph-Petersen, et David A. Lewis. 2005. « Mapping Auditory Core, Lateral Belt, and Parabelt Cortices in the Human Superior Temporal Gyrus ». *The Journal of Comparative Neurology* 491(3):270-89. doi: 10.1002/cne.20702.

Tani, Toshiki, Hiroshi Abe, Taku Hayami, Taku Banno, Naohisa Miyakawa, Naohito Kitamura, Hiromi Mashiko, Noritaka Ichinohe, et Wataru Suzuki. 2018. « Sound Frequency Representation in the Auditory Cortex of the Common Marmoset Visualized Using Optical Intrinsic Signal Imaging ». *eNeuro* 5(2). doi: 10.1523/ENEURO.0078-18.2018.

Taylor, A. M., et D. Reby. 2010. « The Contribution of Source–Filter Theory to Mammal Vocal Communication Research ». *Journal of Zoology* 280(3):221-36. doi: 10.1111/j.1469-7998.2009.00661.x.

Teufel, Christoph, Asif A. Ghazanfar, et Julia Fischer. 2010. « On the relationship between lateralized brain function and orienting asymmetries ». *Behavioral Neuroscience* 124(4):437-45. doi: 10.1037/a0019925.

Tian, B., D. Reser, A. Durham, A. Kustov, et J. P. Rauschecker. 2001. « Functional Specialization in Rhesus Monkey Auditory Cortex ». *Science (New York, N.Y.)* 292(5515):290-93. doi: 10.1126/science.1058911.

Toarmino, Camille R., Cecil C. C. Yen, Daniel Papoti, Nicholas A. Bock, David A. Leopold, Cory T. Miller, et Afonso C. Silva. 2017. « Functional Magnetic Resonance Imaging of Auditory Cortical Fields in Awake Marmosets ». *NeuroImage* 162:86-92. doi: 10.1016/j.neuroimage.2017.08.052.

Tomasello, Michael, et Klaus Zuberbühler. 2002. « Primate vocal and gestural communication ». P. 293-99 in *The cognitive animal: Empirical and theoretical perspectives on animal cognition*. Cambridge, MA, US: MIT Press.

Trapeau, Régis, Etienne Thoret, et Pascal Belin. 2022. « The Temporal Voice Areas Are Not "Just" Speech Areas ». *Frontiers in Neuroscience* 16:1075288. doi: 10.3389/fnins.2022.1075288.

Tsao, Doris Y., Winrich A. Freiwald, Roger B. H. Tootell, et Margaret S. Livingstone. 2006. « A Cortical Region Consisting Entirely of Face-Selective Cells ». *Science (New York, N.Y.)* 311(5761):670-74. doi: 10.1126/science.1119983.

Tsao, Doris Y., et Margaret S. Livingstone. 2008. « Mechanisms of Face Perception ». *Annual Review of Neuroscience* 31(1):411-37. doi: 10.1146/annurev.neuro.30.051606.094238.

Tsapkini, Kyrana, Constantine E. Frangakis, et Argye E. Hillis. 2011. « The function of the left anterior temporal pole: evidence from acute stroke and infarct volume ». *Brain* 134(10):3094-3105. doi: 10.1093/brain/awr050.

Upadhyay, Jaymin, Mathieu Ducros, Tracey A. Knaus, Kristen A. Lindgren, Andrew Silver, Helen Tager-Flusberg, et Dae-Shik Kim. 2007. « Function and Connectivity in Human Primary Auditory Cortex: A Combined fMRI and DTI Study at 3 Tesla ». *Cerebral Cortex (New York, N.Y.: 1991)* 17(10):2420-32. doi: 10.1093/cercor/bhl150.

Vallat, Raphael. 2018. « Pingouin: Statistics in Python ». *Journal of Open Source Software* 3(31):1026. doi: 10.21105/joss.01026.

VanRullen, Rufin, et Leila Reddy. 2019. « Reconstructing Faces from fMRI Patterns Using Deep Generative Neural Networks ». *Communications Biology* 2(1):193. doi: 10.1038/s42003-019-0438-y.

Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, et Pierre-Antoine Manzagol. 2010. « Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion ». *The Journal of Machine Learning Research* 11:3371-3408.

Walker, Jacob, Carl Doersch, Abhinav Gupta, et Martial Hebert. 2016. « An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders ». P. 835-51 in *Computer Vision – ECCV 2016*. Vol. 9911, *Lecture Notes in Computer Science*, édité par B. Leibe, J. Matas, N. Sebe, et M. Welling. Cham: Springer International Publishing.

Walther, Alexander, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, et Jörn Diedrichsen. 2016. « Reliability of Dissimilarity Measures for Multi-Voxel Pattern Analysis ». *NeuroImage* 137:188-200. doi: 10.1016/j.neuroimage.2015.12.012.

Wang, Hongzhi, Yuchao Xu, et Meijing Li. 2011. « Study on the MFCC Similarity-Based Voice Activity Detection Algorithm ». P. 4391-94 in *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*. Deng Feng, China: IEEE.

Wang, Qianshan, Hong Fei, Saddam Naji Abdu Nasher, Xiaoluan Xia, et Haifang Li. 2022. « A Macaque Brain Extraction Model Based on U-Net Combined with Residual Structure ». *Brain Sciences* 12(2):260. doi: 10.3390/brainsci12020260.

Wang, Xiaosha, Yangwen Xu, Yuwei Wang, Yi Zeng, Jiacai Zhang, Zhenhua Ling, et Yanchao Bi. 2018. « Representational Similarity Analysis Reveals Task-Dependent Semantic Influence of the Visual Word Form Area ». *Scientific Reports* 8(1):3047. doi: 10.1038/s41598-018-21062-0.

Welch, Graham F. 1994. « The Assessment of Singing ». *Psychology of Music* 22(1):3-19. doi: 10.1177/0305735694221001.

Welvaert, Marijke, et Yves Rosseel. 2013. « On the Definition of Signal-To-Noise Ratio and Contrast-To-Noise Ratio for fMRI Data ». *PLoS ONE* 8(11). doi: 10.1371/journal.pone.0077089.

Wetzel, Sebastian J. 2017. « Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders ». *Physical Review E* 96(2):022140. doi: 10.1103/PhysRevE.96.022140.

Woods, Kevin J. P., Max H. Siegel, James Traer, et Josh H. McDermott. 2017. « Headphone Screening to Facilitate Web-Based Auditory Experiments ». *Attention, Perception, & Psychophysics* 79(7):2064-72. doi: 10.3758/s13414-017-1361-2.

Wu, Michael C. K., Stephen V. David, et Jack L. Gallant. 2006. « COMPLETE FUNCTIONAL CHARACTERIZATION OF SENSORY NEURONS BY SYSTEM IDENTIFICATION ». *Annual Review of Neuroscience* 29(1):477-505. doi: 10.1146/annurev.neuro.29.051605.113024.

Xing, Chao, Dong Wang, Chao Liu, et Yiye Lin. 2015. « Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation ». P. 1006-11 in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics.

Yamins, Daniel L. K., et James J. DiCarlo. 2016. « Using Goal-Driven Deep Learning Models to Understand Sensory Cortex ». *Nature Neuroscience* 19(3):356-65. doi: 10.1038/nn.4244.

Young, Andrew W., Sascha Frühholz, et Stefan R. Schweinberger. 2020. « Face and Voice Perception: Understanding Commonalities and Differences ». *Trends in Cognitive Sciences* 24(5):398-410. doi: 10.1016/j.tics.2020.02.001.

Zador, Anthony, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, James DiCarlo, Surya Ganguli, Jeff Hawkins, Konrad Körding, Alexei Koulakov, Yann LeCun, Timothy Lillicrap, Adam Marblestone, Bruno Olshausen, Alexandre Pouget, Cristina Savin, Terrence Sejnowski, Eero Simoncelli, Sara Solla, David Sussillo, Andreas S. Tolias, et Doris Tsao. 2023. « Catalyzing Next-Generation Artificial Intelligence through NeuroAI ». *Nature Communications* 14(1):1597. doi: 10.1038/s41467-023-37180-x.

Zäske, Romi, Bashar Awwad Shiekh Hasan, et Pascal Belin. 2017. « It Doesn't Matter What You Say: FMRI Correlates of Voice Learning and Recognition Independent of Speech Content ». *Cortex* 94:100-112. doi: 10.1016/j.cortex.2017.06.005.

Zäske, Romi, Marie-Christin Perlich, et Stefan R. Schweinberger. 2016. « To Hear or Not to Hear: Voice Processing under Visual Load ». *Attention, Perception, & Psychophysics* 78(5):1488-95. doi: 10.3758/s13414-016-1119-2.

Zatorre, R. J., et P. Belin. 2001. « Spectral and Temporal Processing in Human Auditory Cortex ». *Cerebral Cortex (New York, N.Y.: 1991)* 11(10):946-53. doi: 10.1093/cercor/11.10.946.

Zatorre, Robert J., Pascal Belin, et Virginia B. Penhune. 2002. « Structure and Function of Auditory Cortex: Music and Speech ». *Trends in Cognitive Sciences* 6(1):37-46. doi: 10.1016/S1364-6613(00)01816-7.

Zhang, Ya-Jie, Jun-Feng Huang, Neng Gong, Zhen-Hua Ling, et Yu Hu. 2018. « Automatic Detection and Classification of Marmoset Vocalizations Using Deep and Recurrent Neural Networks ». *The Journal of the Acoustical Society of America* 144(1):478-87. doi: 10.1121/1.5047743.

Zhang, Yang, Yue Ding, Juan Huang, Wenjing Zhou, Zhipei Ling, Bo Hong, et Xiaoqin Wang. 2021. « Hierarchical Cortical Networks of "Voice Patches" for Processing Voices in Human Brain ». *Proceedings of the National Academy of Sciences* 118(52):e2113887118. doi: 10.1073/pnas.2113887118.

Zhang, Zhaoyan. 2016. « Mechanics of human voice production and control ». *The Journal of the Acoustical Society of America* 140(4):2614-35. doi: 10.1121/1.4964509.

Zhao, Lingyun, Sabyasachi Roy, et Xiaoqin Wang. 2019. « Rapid Modulations of the Vocal Structure in Marmoset Monkeys ». *Hearing Research* 384:107811. doi: 10.1016/j.heares.2019.107811.

Zhuang, Chengxu, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, et Daniel L. K. Yamins. 2021. « Unsupervised Neural Network Models of the Ventral Visual Stream ». *Proceedings of the National Academy of Sciences* 118(3):e2014196118. doi: 10.1073/pnas.2014196118.

Zoloth, S. R., M. R. Petersen, M. D. Beecher, S. Green, P. Marler, D. B. Moody, et W. Stebbins. 1979. « Species-Specific Perceptual Processing of Vocal Sounds by Monkeys ». *Science (New York, N.Y.)* 204(4395):870-73. doi: 10.1126/science.108805.