UNIVERSITÉ AIX-MARSEILLE

ÉCOLE DOCTORALE 62

Neural, acoustic and information dynamics (mutually) constrain speech perception

Thèse présentée pour obtenir le grade universitaire de docteur en neurosciences par

Jérémy Giroud

Soutenue à AIX-MARSEILLE Université le 17/12/2021 devant le jury:

Dr. Liberty Hamilton	The University of Texas, Austin	Rapportrice
Dr. Jean Luc Schwartz	Université Grenoble Alpes, Grenoble	Rapporteur
Dr. Andrea E. Martin	Max Planck Institute for Psycholinguistics, Nijmegen	Examinatrice
Dr. David Poeppel	Ernst Strüngmann Institute, Frankfurt	Examinateur
Dr. Benjamin Morillon	Aix-Marseille Université, Marseille	Directeur de thèse

À Odile Serdon

AFFIDAVIT

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

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

Fait à Marseille le 11/10/2021

ABSTRACT

Humans are experts at processing speech but how this feat is accomplished remains a major question in cognitive neuroscience. One potential avenue for gaining critical insight into this highly complex cognitive function is to determine the multiple factors that constrain it. During my PhD I adopted two complementary perspectives: a neurophysiological and a cognitive viewpoint to explore speech and language processing.

Specific neural oscillatory dynamics are present in auditory areas during speech perception. These oscillations have been proposed as pivotal processing constraints of the neurocomputational infrastructure, determining the temporal frames of sensory processing. Moreover, due to its evolutionary conserved nature, hemispheric asymmetry has been proposed as a core organizational principle of healthy brain functioning. However, a detailed account of its relevance for speech and language processing remains elusive. In the present work, we provided a fine-grained characterization of the cortical auditory oscillatory dynamics in both hemispheres thanks to intracranial recordings in 96 epileptic patients acquired during pure tones and syllables processing. Our results indicate strikingly similar patterns of activity between left and right hemispheres within the first cortical levels of the auditory pathway. Consistent with the left lateralisation of language functions, a prominent asymmetry emerges in the association auditory cortex. The pattern of activity in the left association auditory cortex is composed of a bimodal distribution of activity (theta/gamma, ~5/40 Hz) compatible with the natural timescales (syllabic, phonemic) of speech.

In the second part of this PhD thesis, we investigated the cognitive constraints at play during spoken language comprehension. Combining a large set of approaches developed to study language comprehension, including linguistics, psychophysics or natural language processing, we developed a unified measurement framework to investigate the respective influence of seven linguistic features on speech comprehension, encompassing acoustic, sub-lexical, lexical and supra-lexical levels of description. Using time compressed speech stimuli within a gating paradigm, we show that comprehension is independently impacted by all these features, but at varying degrees and with a clear dominance of the syllabic rate. We also highlighted the key role of predictive linguistic processes for comprehension.

Taken together the results from the two experimental works shed light on the concurrent neurophysiological, acoustic and linguistic constraints at play during speech processing.

Keywords: neural oscillations, brain asymmetry, auditory cortex, syllabic rhythm, stereo encephalography, speech processing

RÉSUMÉ

Les humains sont experts dans le traitement de la parole. Or les mécanismes sous-tendant cette habileté sont encore mal compris. Étudier les facteurs limitant cette fonction cognitive peut s'avérer être une voie prometteuse pour mieux comprendre les phénomènes complexes impliqués dans le traitement langagier.
Lors de la perception de la parole, on observe des dynamiques oscillatoires neuronales spécifiques dans les régions corticales auditives. Selon des modèles théoriques récents, ces oscillations joueraient un rôle critique dans la perception de la parole. Elles refléteraient des contraintes structurelles biologiques et détermineraient des constantes temporelles pour le traitement sensoriel. Par ailleurs, et en raison de son avantage évolutif, l'asymétrie hémisphérique a été proposée comme un principe organisationnel central du fonctionnement cérébral. Cependant, aucune caractérisation détaillée de son implication lors du traitement de la parole et du langage n'a été réalisée. Dans cette thèse, nous avons caractérisé de manière précise la dynamique oscillatoire dans les deux hémisphères cérébraux grâce à des enregistrements intracrâniens chez 96 patients épileptiques lors de l'écoute de sons purs et de syllabes. Nos résultats indiquent des dynamiques neurales similaires entre les 2 hémisphères dans les régions auditives primaires. Conformément à la latéralisation gauche des fonctions du langage, une asymétrie importante apparaît dans le cortex auditif associatif. Dans la deuxième partie de cette thèse, nous avons étudié les contraintes cognitives en jeu lors de la compréhension de la parole. En combinant un large panel d'approches expérimentales, nous avons développé un cadre de mesure unifié pour étudier l'influence respective de sept facteurs linguistiques à différents niveaux de description sur la compréhension de la parole. Nos résultats indiquent que la compréhension est indépendamment impactée par toutes ces caractéristiques, à des degrés divers et avec une forte dominance du rythme syllabique. Nous avons également mis en évidence le rôle clé des processus linguistiques prédictifs pour la compréhension. En résumé, nos travaux expérimentaux mettent en lumière les contraintes neurophysiologiques, acoustiques et linguistiques simultanées en jeu lors du traitement de la parole.

Mots clés : oscillation, rythme syllabique, parole, cortex auditif, asymétrie hémisphérique, stéréo encéphalographie

ACKNOWLEDGMENTS

The work presented here can in no way be attributed to a single person. It is above all a team effort! In this strange time of global pandemic crisis I think that it is good to remember that science is a collective effort and that it cannot work any other way.

I would like first to acknowledge my supervisor Ben who has been and continues to be an amazing mentor and friend. I have learnt so much from him, and he infected me with his passion, his determination and optimism for scientific questions. I thank him for that.

Thank you Daniele for having created such a stimulating and vibrant scientific and friendly environment within the DCP team, I am fully aware of the chance I have had to be part of the team.

I would also like to thank the members of my PhD committee: Dr. Andrea Martin, Dr. Liberty Hamilton, Dr. Jean Luc Schwartz, and Dr. David Poeppel for reading, evaluating and discussing my work.

This work has been conducted at the interface between fundamental research and the clinical world. The clinical team at the epilepsy unit of La Timone hospital has highly contributed to the development of my neurophysiological knowledge and comprehension. I would like to warmly thank Prof Agnès Trébuchon and Fabrice Bartolomei as well as Patrick Marquis for taking the time to introduce me to the clinical work. I have a great admiration for the nurses in the unit and also the patients that fight everyday trying to have a normal life.

I would like to acknowledge François Pellegrino as well as the researchers in the ‘small’ field of speech and neural oscillations that I had the chance to meet before the covid, for vibrant discussions about science.

Thank you DCP members! This PhD has been fun at your sides: Céline, Noémie, Arnaud, Jacques, Alexia and the others.

I would like to acknowledge the French government, La ligue Française contre l’épilepsie as well as l’ILCB for their financial support without which, the present work would not have been possible.

Finally, a big thank you to my parents as well as my countless brothers and sisters whose joy and support made this scientific journey easier. Last but not least, Lucie, thank you for your unwavering support.

Self disclosure: Although now, I am mostly confident in my ability to correctly localize Heschl’s gyrus, it is still beyond my reach to spell correctly asymmetry in both French and English but I am working on it.

TABLE OF CONTENTS

Part I: General introduction and literature review

1. Speech sounds are (not) of a special kind

Of all the sounds in our environment, the speech signal is probably the most relevant in our daily lives. It naturally and efficiently allows us to establish direct interaction with our peers, to exchange information or to convey emotions. Although it may seem effortless to us, speech processing requires a chain of processing steps that is made even more complex by the fleeting nature of the spoken message and its elaborate spectral content. However, statistical regularities and highly efficient organizational structure of the language system highly contribute to reduce cognitive load during speech processing.

1.1 A complex acoustic signal

i. The Temporal and Spectral Structure of Speech Sounds

Speech results from complex manipulations of airflow through active variations of the vocal tract shape using otolaryngeal structures such as the palate, tongue, and lips (Hickok, 2012).The basic sound, the carrier, is the sound energy produced by the flow of air from the lungs passing through the vocal folds as they open and close rapidly in a periodic or quasi-periodic manner. The resulting carrier is a periodic complex sound with a relatively low fundamental frequency, whose spectrum contains harmonics covering a wide range of frequencies, but with more energy at low frequencies than at high. This spectrum is subsequently modified by the vocal tract which behaves like a complex filter and introduces resonances at certain frequencies (the formants)(Fant, 1971). The formants' center frequencies vary according to the shape of the vocal tract and characterize phonemic categories. The fundamental frequency gives rise to the percept of voice pitch and thus provides an important perceptual cue for speaker gender identification (Klatt & Klatt, 1990; Moore, 2012). Adults males speech typically contains low fundamental frequency values (approx. 75 – 175 Hz) whereas fundamental frequency values are higher for female speakers (175 – 300 Hz) and even more so for infants and children (300 – 500 Hz) (Greenberg & Ainsworth, 2004).

Vowels are the most easily characterized speech sounds as they are pronounced without any stricture in the vocal tract, they are usually voiced and have formants that are relatively stable over time. Consonants are speech sounds whose mode of production is characterised by the constriction of the airflow in the cavities along the vocal tract. The place of articulation (the place where the constriction occurs) and the mode of articulation (the way in which the constriction occurs, partial or total, with or without vibration of the vocal folds, with or without opening of the palate) are reflected in the spectral acoustic characteristics of the speech wave.

On one hand, the spectral or frequency dimension is a crucial aspect of the speech signal; it corresponds to the distribution of the energy of the sound signal in the frequency scale (sound spectrum), and makes it possible to define the different formants of the speech units (in particular the vowels) and their transitions (called formantic transitions characterising the articulation between two speech units; (Stevens & Klatt, 1974). On the other hand, the temporal dimension of the speech sounds is highly relevant for comprehension. This second dimension corresponds to the precise organisation of the different elements of speech over time.

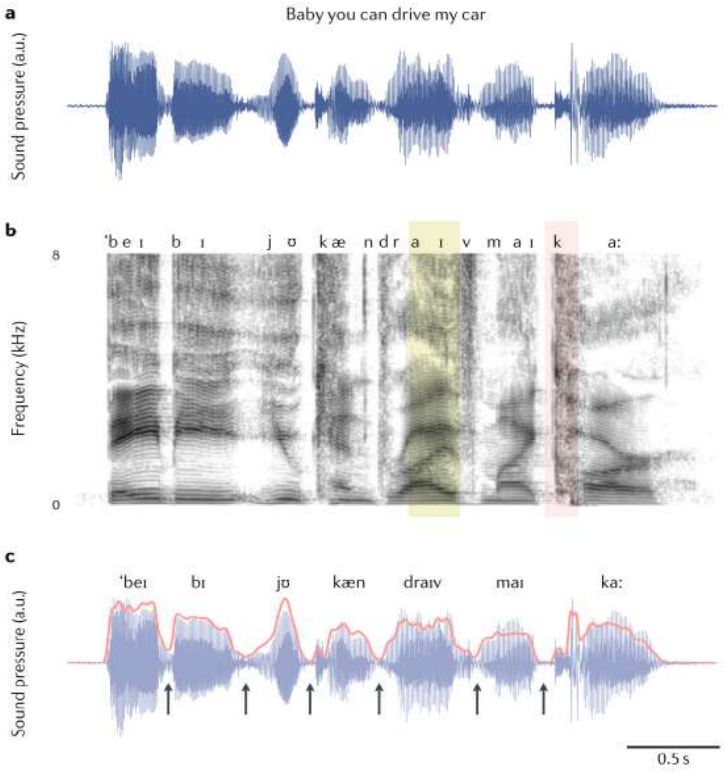


Figure 1. a) Acoustic waveform of the utterance 'Baby you can drive my car' produced by a male English speaker. The y axis represents the amplitude of the sound pressure level in arbitrary units (a.u.); same in panel c. b) Spectrogram of the signal, emphasizing short-scale dynamics. Spectrogram representations allow the visualization of phonemic features over time. For example, dark stripes in the yellow-shadowed region define the /a/ and /I/ vowels, and consonants such as /k/ in the red-shadowed region. c) Mesoscale representation: evolution of the waveform amplitude over time, the so-called speech envelope (pale red trace). Although the envelope — as any other natural signal — is not perfectly periodic, it displays certain temporal regularities. Troughs in the envelope (grey arrows), which typically correspond to syllable boundaries, are roughly regularly distributed in time. ˈbeɪbi juː kæn draɪv maɪ kɑ represents the International Phonetic Alphabet transcription for the corresponding speech chunk immediately below each sign. Adapted from (Poeppel & Assaneo, 2020).

Speech, then, is composed of acoustic patterns that vary in frequency, intensity and time: the result is a modulated signal composed of frequencies between 20Hz and 20 kHz. These highly complex and concomitant amplitude and frequency modulations over time representing low level acoustic features are crucial for intelligibility. Discoveries in the study of speech sounds have largely benefited from useful ways of visualising and representing the signal at different levels of granularities. Various and classical ways of depicting the same acoustic signal are represented in (Figure 1).

Indeed, the speech signal contains several time constants relevant for comprehension. It can be decomposed via the Hilbert transform into three components according to its rate of fluctuations in the time domain: ***envelope*** (2-50 Hz), ***periodicity*** (50-500 Hz), and ***fine structure*** (0.6-10 kHz). Each is associated with several linguistic features categorised into segmental and prosodic information (Rosen, 1992). These spectral and temporal cues relevant to comprehension are redundant in the speech signal.

The temporal envelope also known as amplitude modulations is the lowest frequency component and refers to overall fluctuations in amplitude of the acoustic wave at ~2 – 50 Hz (Figure 1c, red line). The envelope contains a substantial amount of useful speech information, for example, manner of articulation, voicing, vowel identity, and prosody (Rosen, 1992).

The frequency modulations corresponding to the temporal fine structure can be defined as the instantaneous frequency which, regardless of intensity, varies around a center frequency (Moore, 2012) (Figure 1b). The temporal fine structure of speech is comprised between 600 Hz and 10 kHz and is responsible for place of articulation information, vowel quality and segmental cues to voicing and manner (Rosen, 1992).

Speech sounds periodicity gives specific prosodic information related to stress and intonation, as well as segmental information related to voicing and to a lesser extent, information about manner of articulation (Rosen, 1992). Periodic sounds fluctuate mostly at rates between 50 and 500 Hz whereas their aperiodic counterparts fluctuate at a less well-defined frequency range from a few kHz up to 10kHz (Rosen, 1992).

ii. A major feature: the slow amplitude modulations

While a large body of literature has shown a major role of speech spectral information for comprehension (particularly in noise, (Zeng et al., 2005), there has been a paradigm shift in the speech comprehension research field with solid pieces of evidence showing that temporal modulations at low and high rates contained in the speech signal play a critical role for understanding. In particular, there is ample experimental evidence demonstrating strong robustness of our speech perception system to spectrally degraded artificial speech signals, suggesting that our perceptual system can use temporal cues for recognition.

In their seminal study, Shannon and collaborators used vocoded speech (a specific signal processing technique aiming at progressively reducing the speech spectral content) with different numbers of critical bands, thus preserving temporal envelope cues while critically restricting spectral content of the signal (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Their results show that high speech recognition performance can be obtained with only four critical bands of modulated noise. These results indicate that the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for the recognition of speech.

Another piece of evidence in favor of a critical role of the temporal envelope for speech processing comes from clinical settings, where patients with hearing impairments implanted with cochlear implants have a relatively good understanding of speech. This feature is impressive owing to the main limitations of the clinical device: imprecise spectral coding, strong intensity compression and absence of fine structure. Patients' performance in speech comprehension have firmly confirmed the ability of the auditory system to capitalize on the temporal information present in the speech sounds for understanding as the cochlear implant delivers an acoustic signal with a relatively good temporal resolution. Accordingly, these observations implicate that highly impoverished speech signals remain intelligible as long as enough information in the spectrum is available to convey temporal modulations at appropriate rates. Therefore, research agendas have sought to precisely characterize the temporal modulations present in speech and their implication for comprehension.

Because of the way speech is produced (through complex rhythmic and quasi-rhythmic motor gestures involving precise manipulation of the vocal tract articulators), the resulting waveform typically shows a sequence of increments and decrements in signal amplitude in a regular fashion (Figure 1c). A detailed characterization of the temporal modulations present in the acoustic speech signal (which corresponds to its envelope) reveals a highly restricted frequency range of values between 2 and 8 Hz, with a notable peak between 4 and 5 Hz (Ding et al., 2017; Varnet, Ortiz-Barajas, Erra, Gervain, & Lorenzi, 2017) (Figure 2). Critically, this characteristic range is preserved across speakers, languages and speaking conditions (Poeppel & Assaneo, 2020). Speech thus appears to have a regular temporal structure, which the brain might capitalize on to further process relevant information.

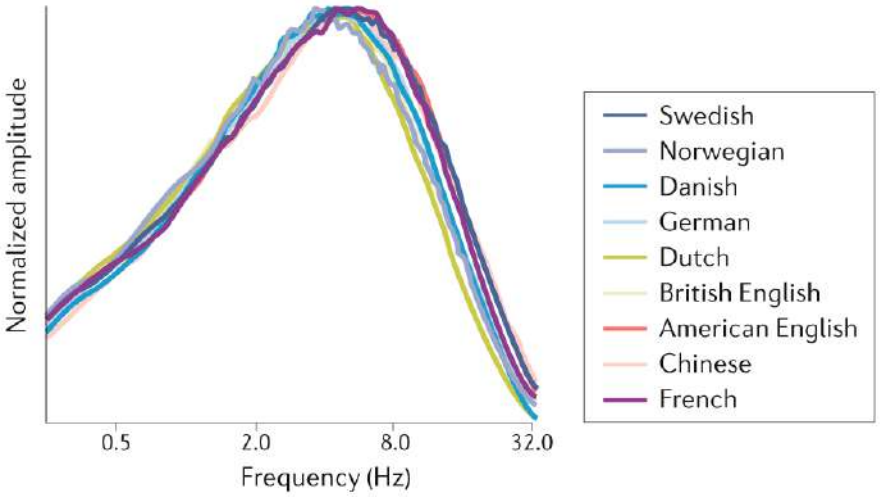


Figure 2. Temporal structure of the speech envelope across languages. Narrowband analysis applied over different speech samples (different speakers, under different speaking conditions including reading, having a conversation or telling a story) from various languages. Independently of the language, the amplitude modulation spectra display clear peaks between 2 and 10 Hz. Adapted from (Poeppel & Assaneo, 2020).

From a linguistic point of view, the syllable is considered a fundamental unit of speech comprehension (Greenberg, Carvey, Hitchcock, & Chang, 2003). Its role is crucial in language acquisition (Mehler & Hayes, 1981). Interestingly, and despite variability in syllable duration, the syllable rate displays a rather restricted range of values. Although significant differences in syllable rate have been reported between languages (Pellegrino, Coupé, & Marsico, 2011), speaking conditions (Künzel, 2013), the variation of this value is always restricted between 2.5 and 8 Hz (corresponding to durations of ~125–400 ms), suggesting that, while speaking, humans sequentially produce syllables at relatively consistent rate which is similar across conditions, speakers and languages.

Rather than being merely incidental, the correspondence between the timescales of acoustic and linguistic units in speech potentially reflect their interconnection, and their critical relevance for speech processing (Poeppel & Assaneo, 2020).

Of course, a large body of evidence points towards a prominent role of the theta timescale (4-8 Hz) for comprehension. A first piece of evidence comes from works that used specific modification of the speech signal. By temporally reversing successive short intervals from a spoken sentence, Saberi and Perrott, showed that when using this specific methodology, participants' intelligibility remains high up to a condition in which 50 ms long reversed segments were used (Saberi & Perrott, 1999). Of relevance, when using spoken stimuli at different speed, the comprehension threshold is not related to the absolute duration

of the time-reverse speech, but rather on the relationship between interval duration and syllable rate: for speech at a higher-speech rate, disruption to speech intelligibility occurs with shorter reversals (Stilp, Kiefte, Alexander, & Kluender, 2010). Hence, comprehension is poorest when syllable-rate intervals are disturbed, regardless of speech rate.

Further demonstrations of the importance of the syllabic time scale in speech comprehension was brought by Ghitza and Greenberg who presented time-compressed versions of spoken materials (reduced to ~30% of the original duration; (Ghitza & Greenberg, 2009). They show that this modification led to impairment in comprehension that was strongly alleviated by the insertion of periods of silence between time-compressed speech segments. In particular, restoring the "syllabicity" of the spoken stimuli (its original temporal structure in terms of the syllable rate), resulted in an optimal identification of the spoken sentences. Several studies further demonstrated that listeners rely on information conveyed in the low-frequency amplitude modulations of the speech signal. In a behavioral study, removing temporal fluctuations that relate to the syllabic rate (2–9 Hz) within spoken stimuli by artificially filtering the signal, resulted in degraded intelligibility for listeners. While artificially reinstoring the temporal modulation by the addition of brief noise bursts that act as temporal cues at exactly where the 'acoustic edges' of the original stimuli would have been, led to a drastic increase in intelligibility (Ghitza, 2012). These results were further confirmed and extended with magnetoencephalographic recordings (Doelling, Arnal, Ghitza, & Poeppel, 2014). Following this line of research, and to overcome the difficulty of defining acoustically the syllable, Ghitza proposed the notion of "***theta-syllable***" a neurolinguistic unit characterized at the interface of linguistic and neurophysiological constraints, inspired by brain function (Ghitza, 2013).

Recently, behavioral investigations of the role of the speech signal isochrony in comprehension in different languages, led Aubanel and Schwartz to the conclusion that the temporal scale associated with the syllable is highly important above its linguistic functional value (Aubanel & Schwartz, 2020). Their results support the usefulness of the notion of the "theta-syllable" that is universally applicable to languages independently of their linguistically-defined rhythm class. Going further, they suggest the P-centre as a potential perceptual candidate as it fits within physiological temporal constraints determined by neural processing in the theta range of 4–8 Hz. Taken together, the above mentioned results are convergent in revealing the pivotal role of the slow temporal modulations present in the speech signal for speech processing.

While slow temporal modulations in speech are linked to the syllabic timescale, higher temporal envelope frequencies are related to segmental information (Rosen, 1992; Shannon et al., 1995). Short-duration cues typically in ~30–50 Hz frequency range are associated with important features of the signal fine structure. For example, at the phonemic timescale (duration of about 20-40 ms), the speech signal envelope conveys information about formant transitions which allow the distinction of different places of articulation crutial for phoneme discrimination (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Furthermore this timescale is also important for relevant speech attributes such as voicing.

In sum, there exists (at least) two different timescales concurrently at play in the speech signal envelope which convey relevant perceptual units for comprehension.

iii. The spectro-temporal modulation framework

Thanks to the spectrogram representation (Figure 1b), it is possible to appreciate the spectral content of the sound over time, a specific processing step that occurs at the peripheral auditory system. Moving beyond this representation, Chi and colleagues presented a biologically inspired model of auditory processing incorporating higher-level processing corresponding to the central auditory system (Chi, Gao, Guyton, Ru, & Shamma, 1999) (Figure 3). In this framework, the modulation domain reflects energy fluctuations that vary across both the temporal and spectral axes of a spectrographic representation and thus describes a specific representational space. Accordingly, the model provides a unified multiresolution representation of the spectral and temporal features of sounds. This model includes two stages of acoustic signal transformation:

First, it is composed of time-frequency analysis based on a cochlear filter bank, then further nonlinear processing steps are applied in order to simulate the characteristics of the auditory nerve fibres to obtain a classic auditory spectrogram. In a nonlinear second step, this auditory spectrogram is analysed by a second, more complex filter bank which models the receptive fields of the primary auditory cortex; the filters are selective to different spectro temporal modulation parameters that range from slow to fast rates temporally, and from narrow to broad scales spectrally. The filters' characteristic time-frequency impulse responses called the spectrotemporal receptive fields are also centered at different frequencies along the tonotopic axis (Chi et al., 1999). A spectrotemporal receptive field is classically described by a Gabor filter with 3 parameters: frequency, time scale and spectral scale. The resulting representation thus has 3 dimensions and is multi-resolution in time and frequency, thanks to the temporal and spectral scales which characterise the modulations of the signal in the two dimensions. The filter output is computed by a convolution of its spectrotemporal receptive field with the input auditory spectrogram which results in a modified spectrogram which gives an index of the modulation energy at different temporal rates and spectral scales. It is therefore possible to highlight the spectro-temporal structures of the sound to which each of the filters are the most sensitive to. Generally, the spectrotemporal receptive fields present excitatory regions, where the sound specific properties lead to energy increases, as well as inhibitory regions that decrease the activation of the filter. Consequently, the filter output is large only if the spectro-temporal modulations are consistent with the rate, scale, and direction of the spectrotemporal receptive field. That is, each filter will respond best to a narrow range of these modulations. This transformation is similar to a two dimensional Fourier transform of the spectrogram, localized at each acoustic frequency (Chi, Ru, & Shamma, 2005).

The modulation domain is an ideal auditory stimulus space for representing speech as the fidelity of the signal encoded by the spectrotemporal receptive field bank has been validated by an intelligibility judgment where listeners obtained nearly 90% of recognised phonemes

when listening to the reconstructed signal (Chi et al., 2005). Moreover, in the perspective of a speech processing model, the multi-resolution representation has a decisive advantage: the different properties of a sound are encoded independently from each other. As a result, the identification of a particular phoneme simply corresponds to the activation of a specific group of spectrotemporal receptive fields. Indeed, studies have shown that individual phonemes are relatively localized in this space (Mesgarani, Cheung, Johnson, & Chang, 2014). Another benefit of this specific representational space is that different types of noise have characteristic distributions in spectrotemporal receptive field space (Chi et al., 2005), to which our speech processing system can capitalize on and hence facilitating speech comprehension in noisy settings.

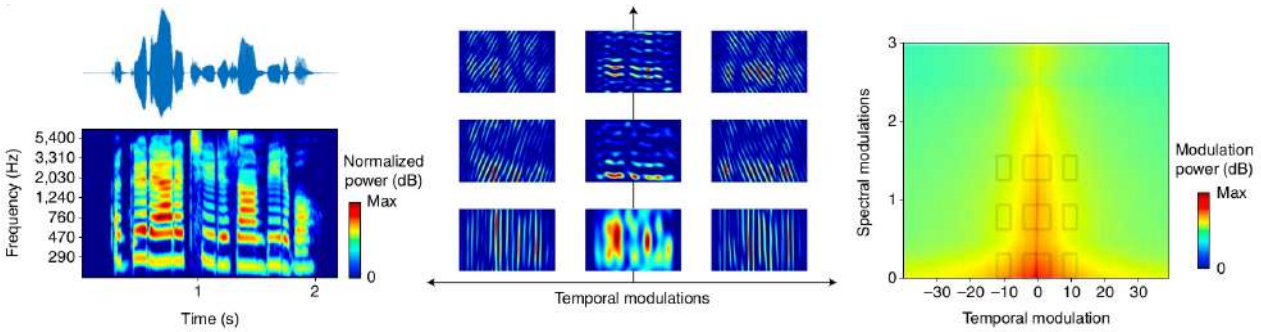


Figure 3. Sound waveform of a spoken sentence (left panel, top) is shown along with its corresponding spectrogram representation (left panel, bottom). The spectrogram can be represented as a decomposition in the modulation domain (middle panel) of horizontal (temporal, typically as cycles per second) and vertical (spectral, typically as cycles per octave) modulations. The degree (power intensity) of temporal and spectral modulations in the spectrogram is depicted in the right panel (showing the average modulation spectra of all the speech material, N=84). Superimposed gray squares correspond to the approximate temporal and spectral ranges shown in the middle panel. Adapted from (Flinker, Doyle, Mehta, Devinsky, & Poeppel, 2019).

1.2 A window into language

i. From continuous stream to discrete representations

The most astonishing achievement of humans is definitely their ability to communicate through language. Language is fundamentally a complex and efficient system giving physical form to meaning using a combination of discrete elements. Using language efficiently (e.g., conveying one's specific message) requires learning the rules that govern it. As such, native speakers of a language have an internal language model which is formed based on linguistic experience and includes knowledge of the likelihood of different words, and the identity of the speech sounds and written tokens that make up those words. However, language is not only a means for humans to achieve shared goals, it is also central to human thought, to social bounding, and to how knowledge is developed, recorded and passed on to next generations. Accordingly, language arises in every human society and while its written and spoken forms are incredibly diverse around the world, they also exhibit surprising concordance (Comrie, 1989). On a daily basis, we most often encounter language through speech. While seemingly effortless and automatic, speech comprehension rests on a complex chain of processes that are both bottom-up and top-down level-dependent (Kuperberg & Jaeger, 2016). Indeed, speech perception proceeds by extraction of sensory information from our external environment (bottom-up processing) in combination with internal information based on prior knowledge, experiences, and expectations (top-down processing) to guide the interpretation of the incoming stimulus and build the final perception. Accordingly there is a dichotomy: whereas speech processing relates to the segmentation of speech into linguistically meaningful units and their identification through the extraction of temporal and spectral cues present in the physical world, language comprehension describes the decoding of the meaning of words and combinations of words, such as phrases and sentences (Meyer, 2018).

The specificity of the speech signal for humans does not simply reside in its complex acoustic characteristics but also in its ability to convey meaning through higher order representations, which can interface with stored representations within the mind of the listener. To comprehend a spoken utterance, listeners must map this highly dynamic and variable spectrotemporally complex continuous acoustic signal onto discrete linguistic representations stored in the brain. The first step of this process relies on an analogical analysis of the signal via the extraction of relevant acoustic cues and possibly the processing of phonetic information. Second, listeners have to assemble these representations in order to extract individual words, access the meanings of these words, and combine them to retrieve the overall meaning (Davis & Johnsrude, 2007). As a result, language comprehension is classically conceived as a hierarchical process which can be broken down into several processing steps, from the low-level extraction of phonetic and prosodic cues, to their

higher-level integration into lexical units and syntactic phrases, and ultimately to global comprehension. It is currently still unclear how listeners proceed from the analysis of speech-specific acoustic structure to linguistic analysis: the so-called acousto-linguistic transformation.

This highly specific human mental set of operations is challenging for several reasons. First and contrary to written texts, natural speech usually comes to the listener as a continuous stream and a detailed inspection of the speech waveform does not reveal clear correlates of what the listener perceives as word boundaries (Klatt, 1980). However, the subjective experience of speech is not continuous but discrete: native speakers of a language perceive a sequence of discrete meaningful units. Consequently, the signal must be first discretized in order to extract relevant linguistic units for comprehension. This parsing operation is rendered challenging as the speech signal is both short-lived and usually unfolds at variable speeds that are not under the control of the listener.

Moreover, another major challenge in speech comprehension results from the fact that multiple talker variabilities such as differences in age, gender, speech rate or accent systematically influence the acoustic regularities present in speech input (Miller, Grosjean, & Lomanto, 1984). Typically, a phone is influenced by the articulation of the neighboring phones because the movement of the speech articulators (tongue, lips, velum) during speech production is continuous and physically constrained. Thus, an articulator may start to move during one phone to get into position in time for the next phone resulting in the phenomenon of coarticulation. There is an analogy with recognizing letters: there are lots of acceptable ways of writing the same character. In other words, there is no one to one correspondence between higher order representations and the acoustic and physical properties of their spoken counterparts. However, human listeners perceive speech in a stable way (the notion of invariance) which reveals the high robustness and adaptability of the underlying perceptual system in its ability to successfully map pertinent acoustic cues into meaningful representations. Some authors even assumed that invariance is not in the signal but in the listener (Liberman & Mattingly, 1985).

The hierarchical structure of language appears to strongly imply the existence of different size units in speech which are combined in different ways to create an infinite number of messages. While currently, there seems to be no consensus on the nature of the fundamental unit of speech recognition, it is generally hypothesized that representations described in linguistics may be at work during language perception. Two levels of phonological description have been subject to intense investigation in the speech sciences: phoneme-sized units (either of a phonetic or a phonemic nature) and syllable-sized units.

Phonemes are the smallest units (segments) of sound in the spoken word and represent a generalization or abstraction over different phonetic realizations. Their duration is between 60 and 150 ms, with the majority being around 50-80 ms (Ghitza & Greenberg, 2009; Rosen, 1992). Phonemes are composed of vowels and consonants and are the smallest perceptual unit capable of determining the meaning of a word (the words beer and peer differ only with respect to their initial phonemes).

Syllables are the phonic units immediately superior to phonemes; their duration is between 150 and 300 ms, with an average around 200 ms (Greenberg, 2001; Rosen, 1992). The syllable is an essential unit in the organisation of speech and language, both with regard to all languages, acquisition, pathologies, language errors and psycholinguistic processing (Dolata, Davis, & Macneilage, 2008). It is difficult to describe in the acoustic domain and some authors have envisaged this entity as being a multimodal representation at the sensory-motor interface (Schwartz, Basirat, Ménard, & Sato, 2012; Strauß & Schwartz, 2017).

Researchers have been captivated by the possible form of the universal perceptual unit of speech. One possible candidate: the phoneme has been put forward following seminal work of Alvin Liberman and his colleagues at Haskins Laboratories in the search for invariants in speech perception which began in the 1950s. Their psychoacoustic works led to the identification of the categorical perception of phonemes and thus resulted in the conclusion that this specific representation was the primary unit of perceptual analysis of speech. Using synthetic speech stimuli distributed at equal intervals along an acoustic continuum leading to perceptually different outcomes, they recorded identification responses from participants. They observed that along these uniform variations of the acoustic signal subjects' identification responses did not evolve in a linear way, but instead seem to mark sharp boundaries: that was the discovery of the ***categorical perception*** (Liberman, Harris, Hoffman, & Griffith, 1957) (Figure 4). Their results implied that not only does categorical perception delineate discrete phonetic regions within the set of speech sounds, but it also distorts the perceptual space according to these categories. The sensitivity of the system to a small physical change in the speech signal is not constant along the continuum: it is perceived very clearly when it straddles a boundary but is blurred when the two stimuli are categorized in the same way. It is as if the listener ignores the acoustic variations within the phoneme categories and focuses only on the acoustic differences between the categories that are relevant for the transmission of the linguistic message. Thus, phonemes with very different acoustic characteristics can be perceptually equivalent.

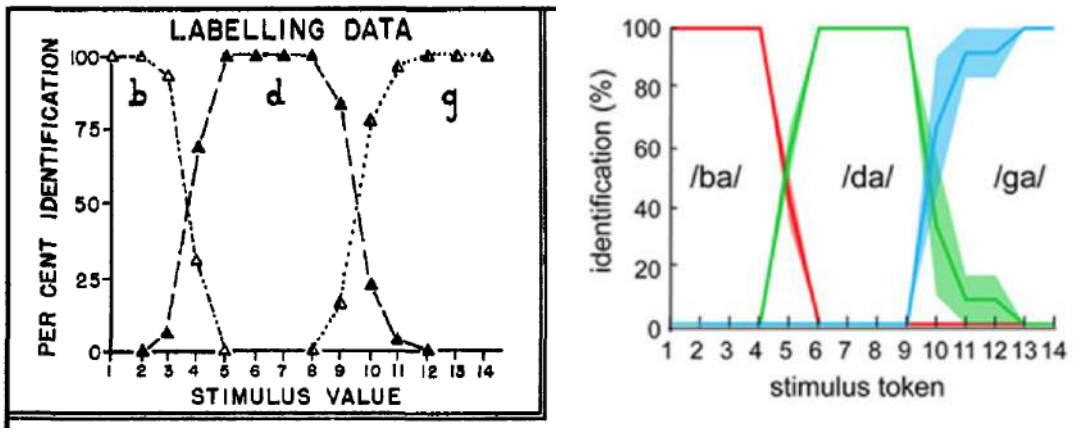


Figure 4. Psychometric identification functions with percentage reporting /b/, /d/, /g/ (left panel) and /ba/, /da/, or /ga/ (right panel) respectively. 14 stimuli were synthesized along a continuum with equal parametric changes in the F2 starting frequency (from 800 to 2100Hz) (right panel). Adapted from (Liberman et al., 1957) and (Chang et al., 2010).

However, as postulated later, phonemic awareness may in fact be a result of recognizing individual words rather than a prerequisite (Charles-Luce, Luce, & Cluff, 1990; Marslen-Wilson & Warren, 1994). Furthermore, speech comprehension is not limited to the identification of a string of phonemes, and in some cases the ability to identify words is preserved while access to the phonemes that make up the word is impaired or impossible (Holt & Lotto, 2010). These findings served to greatly relativize the weight given to phonemes in speech perception. Alternatively, several researchers have argued that listeners make use of alternative representations based on articulatory features (Browman & Goldstein, 1992; King & Taylor, 2000; Stevens, 2002). Articulatory or articulatory-acoustic features describe properties of articulatory events – that is, the lip, mouth and tongue movements that speakers make when producing speech sounds. They are typically described in terms of abstract classes that characterise the most essential aspects of the articulatory properties of speech sounds such as manner and place of articulation, tongue position, and voicing.

Another influential perspective rests on the fact that articulatory features are combined into larger syllabic units during recognition (Greenberg, 1999). It has been proposed that many forms of pronunciation variability can be more effectively modelled using syllables rather than phonemes as the unit of representation (Greenberg et al., 2003). Furthermore, experimental evidence supports the notion that we do not like to segment speech so that it leaves parts of syllables unattached to words: possible-word constraint (Norris, McQueen, Cutler, & Butterfield, 1997). Thus, emphasizing the importance of the syllable as an organizational unit of spoken language. However, there are differences between languages in terms of evidence for the prominence of the syllable as a critical speech unit (Goldinger & Azuma, 2003).

Finally, a new avenue of work in the quest for a canonical unit of representation for speech has emerged. It is based on the recent successes of deep Learning models on a variety of auditory tasks, highlighting their potential relevance for auditory neuroscience research. Within this framework, the model is trained in an unsupervised manner to perform

specific auditory tasks such as sound classification or speaker verification. Looking at the specific learned parameters of the model following the training step completion, has resulted in the successful extraction of interpretable and meaningful representations. Moreover, the retrieved representations were shown to have similar spectro-temporal parameters as the ones measured directly in the human auditory cortex (Riad, Karadayi, Bachoud-Lévi, & Dupoux, 2021). This burgeoning line of research paves the way for the discovery of meaningful representations for speech comprehension without the need for concepts taken from linguistics.

It is still unclear what level(s) of analysis is/are 'elemental' in speech perception. Recent view on the subject emphasizes the fact that speech and language processing are not a monolithic function but depend on a highly dynamic and distributed set of complex processes which take place to some extent sequentially but also in parallel. In this vein, questions remain on the number of canonical levels of representation (e.g., phonemes, syllables, words, etc...), their time constant (e.g., whether they are processed simultaneously or sequentially) and their potential interaction (e.g., larger units (like words) being assembled from smaller elemental units).

ii A complex and hierarchical structure

A detailed examination of the fine-scale organization of the speech signal revealed its multiplexed nature: linguistic information is conveyed over multiple temporal scales (Poeppel, Idsardi, & van Wassenhove, 2008; Rosen, 1992; Stevens, 2000): short energy bursts and formant transitions occur within a 20- to 80- ms timescale, whereas syllabic information occurs over 150- 300 ms and the average durations of words are generally longer. While linguistic processes such as phonetic categorization require the analysis of brief temporal properties in the acoustic signal, speech comprehension ultimately requires the integration over longer temporal windows to extract syntactic, semantic, and lexical information. Moreover, there are correlations which exist over many time scales in the speech signal and result in one of its characteristic attributes; its particular 1/f power spectrum (Voss, 1975).

Speech contains "musical" elements resulting in the so-called prosody which constitute an important low-frequency periodic cue in the speech envelope. Indeed, during speech production, speakers produce their utterances in chunks (intonation units) with a specific prosodic profile. For instance, word endings can be signaled by specific acoustical markers such as lengthening of the last syllable, change in pitch or a decrease in coarticulation between syllables. As such, natural speech is far from being flat; rather, it contains intensity, durational and pitch variations creating rhythms, stress patterns, and prosodic contours. These prosodically defined units aim to pace the flow of information during speech. Sequences of information units form a consistent rhythm of prosodic segmentation cues at ~ 1 Hz (Inbar, Grossman, & Landau, 2020). Speech is organized hierarchically into prosodically coherent units from syllables to sentences and utterances which provide perceptual anchors relieving memory constraints.

Language has been envisaged as the most advanced form of sequence processing for humans as such, it requires different types of processing at multiple levels of analysis. When listening to a sentence, information about the word order is often crucial. Humans then must be able to encode the temporal order of this particular sequence. Another critical aspect of speech is its combinatorial nature, where multiple smaller units can be grouped together to give rise to a new larger unique entity. This chunking process facilitates the manipulation of a group of items as a whole at the next hierarchical level. Combining the information about individual words into an overall representation of the multi word utterance for example, is very frequent during speech comprehension. Usually, speech comprehension depends on the listener's ability to recover hierarchical structures which are the result of the combination of various individual elements into a single higher-level unit. What best characterizes human language is its compositional nature whereby the meaning of a constructed unit is a function of the meanings of its constituents and the way in which these are structurally combined (Partee, 1975, 1995). Language compositionality and hierarchical organization enable unbounded expressive power to speakers. As such, humans can produce and understand in a flexible way words and sentences that have never been encountered before, as long as they respect the language ***syntax***; the rules for arranging items (sounds, words, word parts or phrases) into their possible permissible combinations in a language. Words and phrases can recombine endlessly to form longer sentences, resulting in a never-ending novelty.

Human language is thought to be organized according to nested tree structures generated by symbolic rules. In this vein, a sentence can be parsed thanks to abstract grammatical rules into a set of groupings, which can be embedded within each other resulting in a nested structure of arbitrary depth, and the same elements can be recursively used at several levels (see Dehaene, Meyniel, Wacongne, Wang, & Pallier, 2015 for a review). Sentences can be described as tree-like hierarchical structures of nested phrases which can be moved or replaced as a whole (a pronoun can replace a noun phrase). The power of the postulated tree-like representation of the language lies in the fact that it allows to go further than a simple description of the temporal or ordinal dependency between the word sequence composing a sentence and instead characterize it in terms of structure only at an abstract level. Tree structures are for example indispensable to account for agreement (the link between subject and verb) and binding (the link between a pronoun and its antecedent noun). This specific organizational scheme also is the only one suited to explain Long-distance dependencies phenomena that are typical of languages. Nested structures have another advantage: they permit a compression of the information in sentences and result in an efficient way of coding complexe dependency without overcoming human memory capacity. Some authors have even postulated that the ability to represent tree-like structure is a hallmark of human processing; insisting on their tendency for a dendrophilia (Fitch, 2014).

Speech is inherently a temporal signal which enters the ears in an incremental fashion. Moreover, it is highly transient and fleeting: at a normal speech rate, speakers produce about 10–15 phonemes per second, corresponding to roughly 5–6 syllables every second or 150 words per minute (Studdert-Kennedy, 1986). As such, and given human memory limitations, to understand speech, it is crucial to use all information available at each point in time in an

incremental manner in which representations are built up as rapidly as possible. Furthermore, to correctly retrieve the intended message, processes must take place in parallel across multiple levels of linguistic representation, each with its corresponding characteristic temporal window. Indeed, speech mainly acquires meaning through the combination of its constituent elements extended over time.

In sum, the speech signal is the acoustic medium through which highly structured information is conveyed. Each element is linked to others by more or less complex relationships that allow the construction of organized structures bearing relevant meaning. Humans are expert at effortlessly and rapidly extract such complex structures thanks to potential highly efficient incremental and parallel processing that break free from human memory constraints.

1.3 Speech as a probabilistic signal

i. Initial observations

Given the fleeting nature of speech and human memory limitation, speech comprehension faces the “now or never bottleneck” (Christiansen & Chater, 2016). One way by which our perceptual system can keep up with such a rapid information intake is by predicting what will most likely come next, hence reducing the reliance on solely bottom-up processing. Accordingly, perceptual inferences would be a core principle of brain functions and can be defined as the ability to infer sensory stimuli from predictions based on internal neural representations built through prior experience. In this framework, being exposed to our environment would enable the brain to extract statistical regularities and capitalize on it to create an internal model of the world which would be useful to make predictions in an attempt to more efficiently process sensory inputs. The main goal of this mechanism is to increase the flexibility and robustness of brain processing in highly variable and ambiguous situations.

Ambiguity is ubiquitous in speech comprehension, nevertheless humans are extremely good at extracting meaning from continuous speech. Probability is one of the factors that play a role in disambiguation; the more probable an interpretation, the more likely it is to be chosen. As such, speech comprehension not only rests on simply bottom-up processing but instead relies on the interaction of this sensory source of information with top-down knowledge on the structure of the language at multiple levels.

One classical illustration of this phenomenon is the perceptual restoration, in which a phoneme is perceived within a word despite the fact that the actual phoneme has been replaced by a segment of noise or other nonspeech sound (Warren & Obusek, 1971). An even stronger example of such top-down restoration is illustrated with experimental results obtained using sine wave speech which is constructed of solely pure tones representing only the center frequency and intensity of the formants in the original speech. Perceptual judgments indicate that when naïve listeners encounter these specific stimuli, they report hearing chirps or electronic sounds, on the contrary, when they are explicitly told that they will

hear speech, they are immediately able to hear phonemes (Remez, Rubin, Pisoni, & Carrell, 1981). Such phenomena strikingly illustrate the general principle that human auditory perception is determined as much by what is expected as by what is available from sensory input.

Another clear evidence in favor of a main contribution of probabilistic processing during speech comprehension is the Ganong effect which is characterized by the tendency to perceive an ambiguous speech sound as a phoneme that would complete a real word, rather than completing a nonword (Ganong, 1980). For instance, when presented with exactly the same sound that could be heard either as /g/ or /k/, listeners will likely perceive it as /g/ when followed by “ift” but will perceive it as /k/ when followed by “iss.” Because the target speech sound (/g/ or /k/) is the same across both conditions, this effect demonstrates the influence of top-down processing rather than simply bottom-up processing in speech perception.

ii. Frequency effects

In speech, all words do not have the same probability to be uttered. They obey the Zipfian distribution (Zipf, 1935). This observation has an extremely important consequence on the way they are ultimately processed. For instance, words that frequently occur together can be a reliable source of information for anticipation. Following these early observations, psycholinguistic work has demonstrated a robust effect of the word frequency (the frequency with which a word occurs in natural language) in both comprehension and production. The earliest work has been conducted in the 1950’s by Howes and Solomon, who displayed a word with iteratively longer and longer durations (Howes & Solomon, 1951). They showed that the log frequency of a word (as computed from corpora of over four million words) highly correlated with the mean duration needed to recognize it; recognition of more frequent words necessitated shorter presentations than less frequent words. Later on, the use of the lexical decision paradigm, in which subjects decide if a string of letters presented visually is a word or not, has resulted in shorter durations of lexical decisions about high-frequency words compared to low-frequency words (Balota & Chumbley, 1984; Rubenstein, Garfield, & Millikan, 1970). Similarly robust results have been found for auditory word recognition in which subjects were presented with high and low-frequency words immersed in noise (Howes, 1957). Subjects were better at identifying high-frequency words than low ones. Grosjean developed the gating paradigm, in which subjects hear iteratively more and more acoustic chunks of a spoken word; his results confirmed that high-frequency words were recognized earlier (e.g., given less of the speech waveform) than low-frequency words (Grosjean, 1980). In conclusion, the evidence shows that during speech perception, expectations which stem from statistical properties of linguistic items, act concomitantly to sensory input to determine the content of the spoken utterance; leading to the perception of different phonemes for the same speech sound depending on lexical context, the perception

of phoneme even when such sound is not present in the acoustic waveform, or an asymmetry in the processing of high versus low-frequency words.

iii. Transitional probability

Of note, despite the complexity of structures to be acquired, infants' native language acquisition is quite fast. One of the reasons underlying this success is the fact that the speech signal contains regularities at multiple levels of abstraction. Learners can use statistical learning to perform powerful computations over the distribution of information in a given occurrence, helping them precisely discover the underlying structure of that input. Regularities exist at every level of linguistic structure and can assist the learning process, for example, particular phones co-occurrences can help learners to expand their knowledge about the phonotactics within syllables, at a level above, syllable co-occurrences can constrain the way in which words are composed in the language. Moreover, grammatical and syntactic structure can in part be inferred from regularities in the way words are grouped together in utterances. The role of a distributional learning mechanism has been long shown to drive the acquisition of the ability to segment speech into words. The seminal study by Saffran, Aslin, & Newport, is one of the first experimental works demonstrating the sensitivity of children to transitional probabilities between syllables in speech (Saffran, Aslin, & Newport, 1996). Efficiently using the transitional probability variation present in speech can assist its segmentation since it is low for words, compared to that between syllables within words, which is often much higher. Saffran and colleagues demonstrated that 8-month-old infants were able to follow co-occurrences between adjacent syllables in speech, and distinguish syllable transitions that had different co-occurrence probabilities (Saffran et al., 1996). In the study, a continuous stream of highly constrained artificial speech sounds was used which contained four trisyllabic words. The words were composed of specific triplets of syllables which always occur together. Finally, each word could be followed by any of the other three words. As such, within words, the probabilities of transition between syllables were 1, while the probabilities of transition between words were much lower (0.33). Following a listening session, children gaze time on either language words or on trisyllabic sounds that spanned two words (partial words) were recorded to see whether learners were sensitive to distributional of the language they just heard. Results indicated that children looked longer at partial words compared to words, indicating a preference for novelty in non-word sequences that contained lower transition probabilities. This experiment thus demonstrated a possible role for statistical learning in speech segmentation. Similar results have been obtained in subsequent research, both in children and adults (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997).

Moreover, systematicity in spoken language (e.g., a statistical relationship between the patterns of sound for a group of words and their usage) may facilitate speech processing. For instance, given exposure to enough words, subtle statistical differences in word forms (e.g., stress, duration, voicing) can help listeners to identify grammatical categories (Cassidy & Kelly, 1991). Further works have demonstrated a link between distributional statistical information present in speech and grammatical categories of words. Mintz found that a small number of frequent frames (frequently co-occurring words separated in speech by one other

word) could give robust information about word grammatical categories that occur within the frame (Mintz, 2003). For example, nouns can occur in the frame “the __ is” but verbs cannot, alternatively, verbs can occur within “he ___ the” but nouns cannot. Listeners can then use simple distributional information within spoken language as heuristic to determine the grammatical roles of words (nouns are frequently preceded by articles: the, a).

Humans show a striking ability to understand continuous speech even under adverse listening conditions. The mapping between the time varying signal into discrete representations must then occur nearly instantaneously. One potential mechanism facilitating this process rests on the ability of the listener to constantly integrate incoming sensory information and prior knowledge to produce a unified perceptual experience. As such, following exposure to speech throughout life, speakers acquire statistical knowledge on their native language. These distributional information at multiple levels of linguistic description act as prior knowledge and serve as prediction to constrain potential interpretation of the spoken utterance. As a result, the brain makes use of top-down information to complement bottom-up sensory information as a means of facilitating perception and enhancing speech tracking.

iv. Contextual effects

In everyday life, phonemes, syllables or words are rarely presented in isolation. Instead, they are part of larger and more complex structures such as phrases, sentences or even stories. As such, the context in which isolated units appear potentially gives useful and important information at multiple levels, which the brain can capitalize on to make predictions. Such mechanisms can in turn facilitate and speed up the processing steps required for comprehension. This process critically relies on a dynamical update of the predictions of incoming sensory information, based on internal models acquired through past experiences with language and speech.

During speech, it is often the case that the acoustic properties of the spoken utterances imperfectly match the long-term speech regularities that have shaped the mapping of speech acoustics to linguistically significant representations like phonemes and words. However, and despite challenges imposed by talker differences such as dialects, accents, and other factors that systematically influence the acoustic regularities present in speech input, speech comprehension remains robust in various contexts. Multiple empirical findings have demonstrated that listeners make use of specific acoustic contextual information to adaptively and predictively process speech in this specific situation (Idemaru & Holt, 2011; Zhang, Wu, & Holt, 2021).

Furthermore, in speech communication, not only the nature but also the timing of events is highly relevant for comprehension. As an illustration, Kösem and colleagues had participants listen to sentences with varying speech rates in which the beginning of the sentence was either time compressed or expanded, leading to a fast or a slow speech rate (Kösem et al., 2018). The final three words were presented at the original speech rate. Participants were asked to report their perception of the last word of the sentence, which contained an ambiguous vowel that could lead to the perception of two distinct words. The

study was conducted in Dutch where a similar vowel at the end of a word can change the meaning of that word depending on its duration. Results indicate that the preceding speech rate affected the perception of the target word with participants being more biased towards the perception of the word with a long vowel after a fast speech rate and towards the perception of the word with a short vowel after a slow speech rate. These results are in line with previous experimental works in which contextual speech rate has been shown to affect the detection of subsequent words (Dilley & Pitt, 2010), word segmentation boundaries (Reinisch, Jesse, & McQueen, 2011), and perceived constituent durations (Bosker, 2017).

There is now extensive evidence that prediction is important for language comprehension, just as it is for perception and cognition more generally. Of course, listeners interpret language extremely rapidly. They process each word as they encounter it and integrate it with prior context in a highly incremental manner. They may not only analyze each word as they encounter it, but also predict what they are going to encounter.

Classically, in experimental settings, predictability has been assessed by asking participants to complete a sentence context, using the so-called Cloze procedure (Taylor, 1953). Participants are shown sentence contexts with an empty space instead of the target word. Then, they are asked to fill the blank with a plausible word to complete the sentence. In the case where most participants give the same completion, the context is considered as predictive or constraining and this completion highly predictable or high-Cloze, alternatively, in the case where variable responses are given by the participants, the completion is low-Cloze. In their seminal study, Kutas and Hillyard, demonstrated that sentences with less expected endings result in a larger negative deflection of the electroencephalographic (EEG) signal 400 ms after the onset on the closing word: the classical N400 component (Kutas & Hillyard, 1984) (Figure 5). The relationship between the brain signal and the predictability is very strong and has been reproduced many times. However, two particular facts merit consideration: first, in naturally occurring contexts, usually words are rarely predictable and contexts are most likely medium to low-Cloze, as such some authors have argued that predictions appeared to have a very limited value (Stanovich & West, 1979). Moreover, N400 effects are also compatible with an integration account in which unsurprising words given the context are more easily integrated compared to highly surprising words. As a result, the N400 component is smaller for the former type of words compared to the latter. While there is still a vibrant debate on the real nature of processing implicated in speech comprehension, recent accounts view prediction as a graded and probabilistic phenomenon. This view is based on strong evidence of graded effects of context on processing. For example, the magnitude of the N400 effect evoked by an incoming word is inversely correlated with that word's Cloze probability (DeLong, Urbach, & Kutas, 2005). Experimental evidence has accumulated over the decades and has shown that listeners can use information in a context to facilitate the processing of new information at multiple levels of representation, and that multiple types of information from internal representation of context can be used to facilitate speech processing. For instance, contextual information can be efficiently used to predict the semantic categories, the syntactic structure, and phonological information of the forthcoming speech input (Kuperberg & Jaeger, 2016).

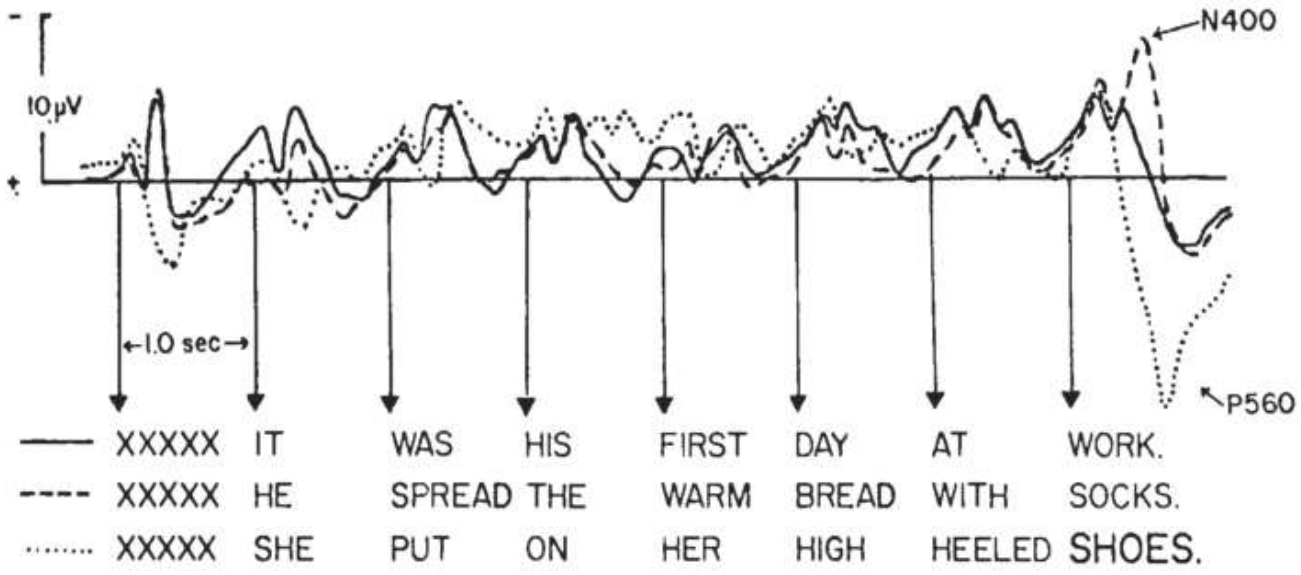


Figure 5. Event related potentials (top) averaged across three subjects and recorded in response to seven-word sentences (bottom). Solid line represents data from the congruous condition, resulting in a low amplitude N400. Dotted-line relates to the incongruous condition and exhibiting a large amplitude N400.The solid and dotted-line conditions are negative-going waves (peaks up). The micro-dotted line is the control condition. It shows a positive-going wave (peak down) that relates with the size of the font of the written stimulus and not with its semantic congruence. Adapted from (Kutas & Hillyard, 1980).

1.4 The Humans' favorite information transmission channel

i. From speech rate to information rate

Speech is one of the preferential means by which humans exchange information. The purpose of speech, then, is not to exchange auditory signals but to exchange information content. Interactions between a speaker and a listener can be formally described as a communication system (Shannon, 1948), where the aim of the speaker is to conceive of a conceptual message and encode it into physical auditory signals. The role of the listener is to decode the intended message from the acoustic signals through a set of complex computations happening at the subcortical and cortical level. The communication can be considered as successful when the reconstructed conceptual message within the mind of the listener matches the intended message of the speaker. The transmission of information proceeds through a channel with a specific capacity which is constrained by specific characteristics of the speakers, hearers, and environments in which this communication takes place. Strategies that take into account these various sources of limiting factors must be applied in order to efficiently communicate.

To process speech, listeners can make use of two important properties of the language. First, while speakers can convey an endless amount of information as long as they play by the rules (e.g., constrained by the vocabulary and grammar), not all expressions are equally probable: some syllables sequences, words and meanings are much more likely than others. This difference in likelihood is not directly visible in the sensory signal, but comes from knowledge based on internal models of the language which include the statistical structure (what items are more or less likely) as well as linguistic regularity (what items are permitted). During speech perception, listeners make use of the inherent statistical regularities present in their language. Such a source of information is critical for comprehension. Information content of speech signals at all levels of the linguistic hierarchy, covering all units between sounds and meanings can be quantitatively estimated within the Information theory framework (Shannon, 1948). Within this framework, information does not refer to the semantic content of the actual message but rather it refers to its unpredictability or unexpectedness. Thus, a message is informative if it is not predictable or expected and conveys something surprising and new.

The surprisal of an item (phoneme, syllable, word) is the log inverse of its probability, that is the probability that the item will occur, hence it represents the item unpredictability. When the logarithm is taken in base k = 2, information content is measured in bits. As such, surprisal is low for highly predictable sequences. Surprisal is an estimate of both the degree to which a particular event is predicted, and the information gained by experiencing that particular event; both are measured in bits. If an outcome is more likely, it is more predictable,

and the information gain is smaller than if the outcome was less likely. If an event outcome has a probability of 1, predictions are entirely confident and no information is gained.

$$Surprisal \quad h(x) \; = log_k \frac{1}{p(x)}$$

The entropy refers to the uncertainty about the outcome of event *x*: it is expressed in bits. A higher degree of uncertainty will result in higher entropy values (due to a higher number of possible candidate continuations) whereas lower entropy values indicate a higher degree of certainty with fewer, highly probable continuations given the context encountered so far. Mathematically, entropy at the current item position H(t) is defined as the expected value of surprisal for the upcoming item (i t+1) given the items encountered so far (i 1 , ..., i t).

$$Entropy \quad H[X] \; = \sum_x p(x) \; log \frac{1}{p(x)}$$

It is possible to compute these information-theoretic metrics at different levels of the linguistic hierarchy and also to adjust the size of the context to take into account to have a relevant description of the speech input given the specific situation under study. In experimental settings such information-theoretic measures are thought to operationalize hypothesized cognitive load or processing difficulty associated with processing individual items in sentences (Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016). There is now experimental support for the sensitivity of the brain to such properties of the speech signal (Donhauser & Baillet, 2020; Frank, Otten, Galli, & Vigliocco, 2015; Schrimpf et al., 2020).

Thanks to information-theoretic measurement it is possible to have a formal description of the informational content of speech signals in the sense of Shannon at any level of linguistic unit. As speech is inherently a temporal signal, previous studies have sought to characterize in great detail the capacity of the speech signal to encode and transmit information over time. As such, Pellegrino et colleagues, took a cross linguistic approach and compared multiple languages on their propensity to convey information (Coupé, Oh, Dediu, & Pellegrino, 2019; Pellegrino et al., 2011). They reasoned that language from all around the world probably would have been shaped by several factors inherent to human cognitive architecture, thus resulting in specific and similar trends in their nature or structure. Specifically, their work was based on different hypotheses. One of these, is the uniform information density hypothesis which states that due to channel capacity, speakers modulate the information density of their utterances in order to optimally transmit the information at a uniform rate, near the channel capacity (Levy & Jaeger, 2007). The ***channel capacity*** is defined as the maximum rate at which information can be transmitted. Another related hypothesis is the equal overall communicative capacity of languages postulating that despite large variation in encoding strategy and linguistic complexity, languages transmit the information at a relatively similar rate within a limited range of variation. As such, a trade-off

exists between speech rate and information density as to render the communication as efficient as possible. First, they determined in a large number of languages each specific ***speech rate***: the average number of syllables uttered per second. Then they computed in a language dependent manner, the ***Information density*** which can be defined as the average density of information per linguistic unit (relative to the Vietnamese language and taking the syllable as unit of reference). Finally, they could retrieve for each language its ***information rate***: the average amount of information transmitted per second. Their results indicate that the encoding capacity of multiple languages as exemplified by their information rates is quite similar and gravitates around 39 bits per second (Coupé et al., 2019). This result points towards a complex and adaptive strategy of various languages which results in an overall similar information transmission capacity. This is a non-trivial phenomenon, when considering that each language is constrained by multiple factors including sociolinguistic, cognitive but also factors related to the capacity of human audio channel.

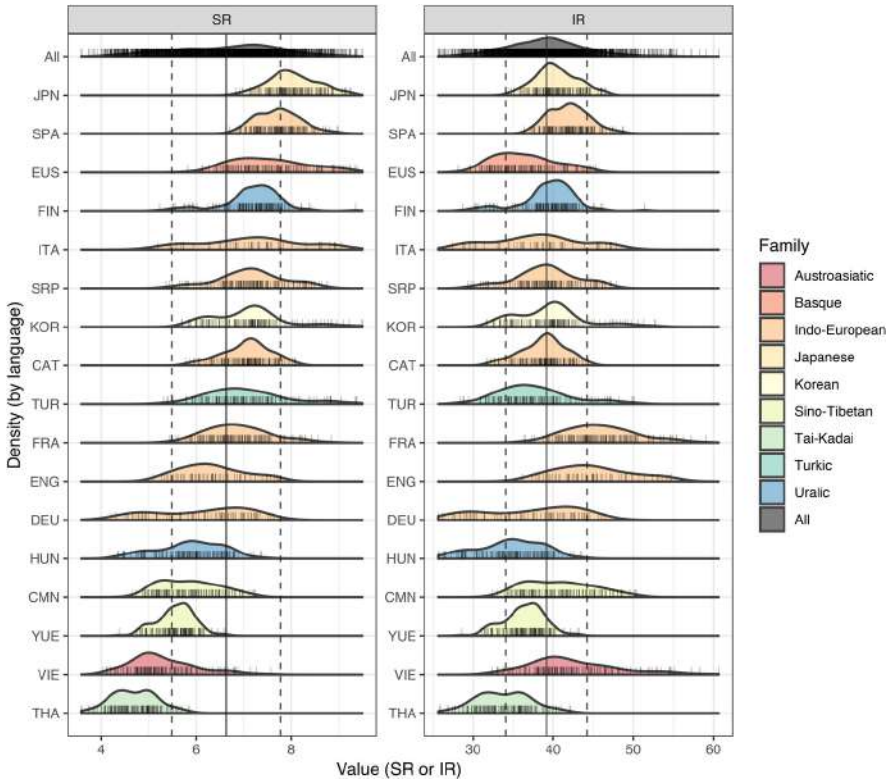


Figure 6. The distribution of syllabic rate (SR; in syllables per second) (left) and information rate (IR; in bits per second) (right) within the languages in the database (colored areas; colors represent the language families) and across them (black areas at the top) using a Gaussian kernel density estimate. The black vertical lines spanning the whole plot represent the means (solid lines) ± 1 SD (dashed lines). The short black vertical lines represent the actual data points. Adapted from (Coupé et al., 2019).

ii. The functional load

The way a language is organized and structured will ultimately have critical consequences for its transmission and comprehension by its speakers. One important factor relates to the confusability and distinction between words within a specific language. Indeed, speakers and listeners must be able to understand each other without too much effort. The phonological components of language then play a major role in distinguishing among its numerous words. While there are different strategies in the way languages convey information, the idea of a detailed description of the implication of phonological structure of the language in this endeavor is not new. The notion of ***functional load*** quantifies the role a phonological contrast plays in keeping words distinct in a given language (Hockett, 1955). As such it is a tool for measuring the relative importance of phonological contrasts in keeping the utterances of a language apart. It has been proposed that some contrasts play a more important role than others in the lexical access and in morphological strategies (Hockett, 1966). For example, some phonological contrasts in French differentiate hundreds of word pairs while others may only be involved in a handful of word pairs. The uneven distribution of functional load is thought to originate from the self-organization of the phonological system as an adaptation to be more resilient and robust to errors and damages of components during speech communication (Beckner et al., 2009). It is possible to select phonological units at any level and consider them as the elementary bricks on which contrasts between words are built. The most obvious procedure to identify them is by listing minimal pairs (when they exist): two sound sequences associated with two different meanings and differing by only one element. The set of such 'distinctive' elements constitutes the phonemic system of a particular language at the selective level of analysis. The functional load is then used for measuring the relative importance carried by any level of granularity of phonological units contrasts and it corresponds to the change of Shannon entropy of the phonological system if the contrasting pair is merged into one unit. For instance, consider the following French words: "raisin" (grapes, /RE/-/z5/) and "raison" (reason, /RE/-/z§/) in the lexicon which is of finite size of all possible French words, the selected pair is only distinguishable by the contrast between /z5/ and /z§/ at the syllabic level. Upon merger of this contrast, rai*** and rai*** would not be dissociable from one another, leading to a decrease in the total number of words in this new fictive lexicon. Hence neutralizing this contrast will make fewer words distinguishable. This will result in an increase of homophony and in changes in the distribution of word frequencies. As the entropy is computed on the whole lexicon and involves probability estimations, the merging will modify its value. The functional load of a contrast between the two syllables /z5/ and /z§/ is defined as the relative difference of entropy between two states of language: the observed state (L) and a fictional state ($L^*_{z5,z§}$) in which the contrast is neutralized and is expressed as a ratio theoretically ranging from 0% to 100%.

$$Functional\ load\ \ FL\,(x,y)\ = \frac{H(L) - H(L^{*}_{xy})}{H(L)}$$

As a consequence, the functional load of the /z5/-/z§/ contrast both requires a global knowledge of the lexicon and measures the impact of the /z5/-/z§/ contrast on the whole lexicon. In addition, one can also focus on the level of syllables themselves, by summing functional load over all the contrasts in which the syllable /z5/ is involved. The functional load of /z5/ thus measures the importance of the syllable /z5/ in the language lexical network. Looking at the syllabic level, Pellegrino and colleagues found that a lot of contrasts were characterized by a very low functional load ultimately conveying a very low amount of information (Oh, Coupé, Marsico, & Pellegrino, 2015). They interpret this finding as the consequences of cognitive efficiency and optimization for language acquisition and information retrieval. As such, language specific phonological properties may emerge from a set of nonlinguistic abilities including cognitive, motor, perceptual and communicative considerations. In other words, languages have been structured by their usage to optimally encode and transmit information in human communication. This fact is thus reflected in their phonological properties (a few contrasts conveying a high amount of information).

iii. An efficient and robust code

In quantitative linguistics and psycholinguistics, it has been argued that human languages are structured for optimal and efficient communication (Frank & Jaeger, 2008; Levy & Jaeger, 2007). Within the previously introduced information-theoretic framework, during a communication, the speaker is the information source. She sends a message through a spoken utterance to the listener who is the receiver and destination. In this situation, human languages are codes that permit information to be transmitted through the audio channel. Communication is successful when the message at destination matches the intended message at the source. However, usually, communication happens within non optimal settings, and there could be noise or error along the transmission process. To avoid such flaws in the communication, robust code must be preferred. It is starting to be acknowledged that due to a pervasive pressure for efficiency, the forms taken by natural languages can be described as such a type of code.

Communicative efficiency relates to the fact that successful communication can be achieved with minimal effort by both the sender and receiver, usually by using short messages which also maximize the rate of communicative success.

Thanks to methodological and technological advances which now enable more precise investigations of language properties at various levels of granularity, a growing body of evidence suggests that despite large differences, languages have been shaped for

communicative efficiency. One of the most acknowledged pieces of evidence in favor of the efficient nature of language comes from the well-known statistical law of linguistics, highlighted by Zipf and already introduced in a previous part of this dissertation. First, he observed that the most frequent words are used much more often than the less frequent words: following a Zipfian rank-frequency distribution. Some authors have argued that this particular organization of the lexicon enables more efficient communication and may arise from pressures towards optimal communication and a minimization of effort by both speakers and listeners (Ferrer-i-Cancho, 2016). In addition to the observation of the peculiar distribution of word lengths within the lexicon, Zipf also observed that more frequent words tend to be shorter, revealing an intrinsic link between word usage and their form. The reason for the correlation between length and frequency is that short words are less costly in terms of cognitive load as such they should be preferred and thus used more often than longer words. Zipf called this phenomenon the Principle of Least Effort and is often referred to as Zipf's Law of Abbreviation. Following information theory, if one wants to construct an optimal code being as short as possible while still being robust, word lengths should be proportional not only to frequency but rather to predictability in context. Indeed and consistent with this account, it has been repeatedly observed that a word's probability within a given context was a stronger determinant of its length (number of letters) than its sole frequency of occurrence (Piantadosi, Tily, & Gibson, 2011).

More examples coming from specificities of multiple language lexicons are in line with the principle of communicative efficiency. An efficient communication rests on a tradeoff between two opposing factors. On one hand, the concept of simplicity which in the examined case refers to the ability of a communication system to enable information transmission and reception with compact cognitive representations. One extreme example of this property would be a communication system which contains a unique word for all objects in a given semantic system. As can be imagined, such a system would not support precise communication. As a result, the informativeness of a system is the other driving force in its general organisation. It is related to the precision offered by the means of communication. A maximally informative system would have a unique and separate word for each object in a given semantic domain resulting in high cognitive and memory load. Experimental evidence supports the idea that the general semantic organization of the world's languages optimally balances simplicity and informativeness: enabling highly informative communication at the cost of only modest cognitive complexity. This specific trade-off has been described in various languages within specialized semantic spaces such as color words, kinship terms, number and naming systems. For example, languages such as English or Spanish which are used in industrial nations have been found to have significantly more information in their color system than languages from isolated Amazon groups such as the Tsimane' language. The hypothesized reason for this difference is that the Tsimane' have less need for a highly informative color system since they are less likely to differentiate objects based on color due to the environment in which they live. Therefore, color words seem to be adapted to the needs of the speakers illustrating the concept of efficiency.

Another example is related to the well-known phenomenon called the arbitrariness of the sign. It refers to the fact that word forms are not systematically related to their meaning and it is a well-established property of the world's lexicons (Hockett, 1960). However, multiple experimental observations have revealed lots of exceptions. For example, iconicity (Perniss, Thompson, & Vigliocco, 2010) which links acoustic properties of words to aspects of their meanings. One line of argumentation rests on the so-called "bouba-kiki" effect, where cross-linguistically, people tend to give rounded and smooth objects names with labial consonants and open vowels ("bouba"), whereas spiky, sharp objects are more likely to have sounds with closed vowels ("kiki"). Another phenomenon is systematicity and refers to the fact that words show systematicity in statistical regularities between their forms and their usage within a specific language. For instance, grammatical classes, such as nouns or verbs share certain phonological and prosodic properties. Studies have revealed a systematic positive correlation between the phonological similarity of word forms such as the number of phonemes they share and the semantic similarity of their meanings, this link is significantly above what would be expected under random form-meaning assignment (reviewed in Gibson et al., 2019). These multiple observed phenomena are thought to be the result of a trade-off between simplicity and informativeness in order to ultimately serve the final goal of the human communication system in its purpose of conveying information in an efficient manner. Whether almost all experimental support in favor of such a goal has been described here at the level of the lexicon, efficient design properties at other levels of granularity has been put forward as for instance the organization of the syntax to optimize language efficiency. A better understanding of the cognitive and communicative constraints at play during communication will undoubtedly lead to better characterization of the organisational properties of the language system. Recently some authors have advocated for a larger investigation of similar phenomena (linguistic laws) applied to multiple fields of biological sciences (biological laws) (Semple, Ferrer-i-Cancho, & Gustison, 2021).

2. The cortical organisation of speech processing

The temporal resolution of the auditory system is considered at least an order of magnitude higher than the resolution of any other sensory system (Griffiths, Uppenkamp, Johnsrude, Josephs, & Patterson, 2001). It is amazing how simple movements of molecules in the air can give rise to complex conceptual representations in the brain. The specifics of the mechanisms and neural bases underlying this incredible ability are slowly beginning to be identified and apprehended. From Marc Dax's first descriptions to the ever more precise characterization of the complexity of the neural architecture supporting speech perception, we are closer than ever to having a complete picture of the capacity that truly makes us human. With the increasing temporal and spatial resolution offered by neuroimaging techniques, we are moving away from the static and highly lateralized historical view of speech processing. The current framework postulates that speech understanding is supported by a more dynamic set of processes that are distributed across both hemispheres. However, there is still a long road ahead towards a computational(ist) neurobiology of language. One of the remaining puzzles, of great interest to this dissertation, concerns the partially bilateral and partially lateralized brain representation of speech and language and its relation to other natural sounds.

2.1 theoretical perspectives

i. Is speech processing in the brain "special" ?

Within the auditory neuroscience community, there exist contrasting views about the extent to which the processing of music, speech, and other natural sounds relies on shared or distinct neural mechanisms. The domain-general view predicts that speech (like other natural sounds) processing is supported by general-purpose mechanisms sensitive to the low-level acoustic features present in the signal (Albouy, Benjamin, Morillon, & Zatorre, 2020; Poeppel, 2003; Schön et al., 2010). In contrast, the domain-specific view postulates that speech is processed in a dedicated system which critically depends on the specific linguistic properties of the stimulus (Finkl et al., 2020; McGettigan & Scott, 2012).

While the debate is still ongoing at present, partial but encouraging answers can be provided when envisaging the question within specific conceptual frameworks such as the efficient (neural) coding hypothesis. This principle formulated by Barlow (Barlow, 1961) argues that our perceptual systems have evolved to encode ethologically relevant stimuli in our environment in the most efficient way. This is best achieved if the statistical structure of natural stimuli along the relevant dimensions are neurally encoded in a sparse, non-redundant (statistically independent) and optimal manner. This theoretical framework can

be precisely defined in mathematical terms within information theory; thus, making it empirically testable. As such, an efficient code is defined as one that transmits the highest fidelity information contained in the signal (maximizing the information conveyed) at the lowest cost for the sensory system (minimizing the required energy and neural resources). The codes used by different organisms might be adapted to their different and respective natural auditory environments but also shaped throughout evolution and development (Figure 7A-B).

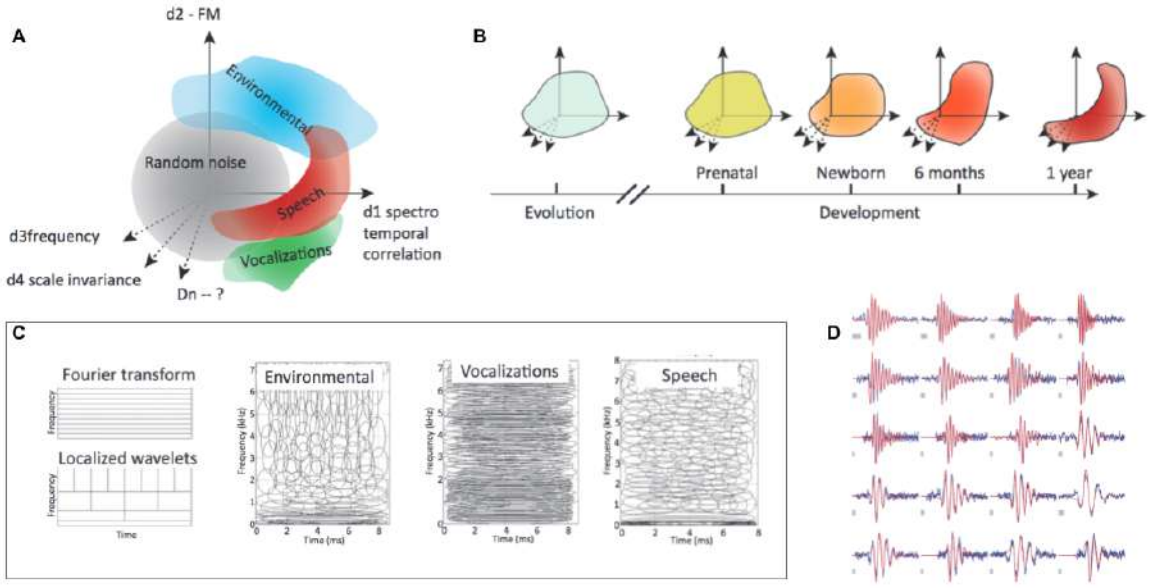


Figure 7. A) Diagram of the spectro-temporal statistical space of different types of sounds projected on a subset of dimensions, including d1, Spectro-temporal correlation; d2, frequency modulation (FM); d3, Frequency; d4, scale-invariant coefficient; dn -- other components to be identified. B) Speech statistics are shaped throughout evolution and development. C) The spectro-temporal characteristics of mathematically computed optimal filters. The filters in a Fourier transform are localized in frequency but not in time (top left). Wavelet filters are localized in both time and frequency (bottom left). The statistical structure of the signals determines how the filter shapes derived from efficient coding of the different data ensembles are distributed in time–frequency space. Each ellipse is a schematic of the extent of a single filter in time–frequency space for environmental sounds, animal vocalizations and speech, second, third and last panel respectively. D) Kernel functions optimized to encode an ensemble of speech sounds. Each kernel function (red curves) is overlaid on a revcor (reverse correlation) function obtained from cat auditory nerve fibres (noisy curves in blue), note the close match between them. A, B adapted from (Gervain & Geffen, 2019). C adapted from (Lewicki, 2002). D adapted from (Smith & Lewicki, 2006).

Natural sounds span a specific range of all possible sounds: their joint temporal and spectral modulation spectrum occupy a restricted space: sounds cannot have rapid temporal and spectral modulations simultaneously. Within this restricted space, they have a characteristic signature (Figure 7A). Natural sounds, in general, are low-passed, showing most of their modulation energy for low temporal and spectral modulations. Animal

vocalizations and human speech are further characterized by the fact that most of the spectral modulation power is found only for low temporal modulation (Singh & Theunissen, 2003). Based on the hypothesis that speech perception may rely on the same auditory coding mechanisms that facilitate efficient coding of other natural sound statistics, Lewicki used independent component analysis to derive efficient codes for different classes of natural sounds, including animal vocalizations, environmental sounds and human speech (Lewicki, 2002). He was able to retrieve theoretically optimal codes as represented by the properties of recovered filters (Figure 7C). He showed that the optimal filters for animal vocalizations resembled a Fourier decomposition (Figure 7C). It was not surprising given the high prevalence of harmonic structure in these sounds; the filters for environmental sounds looked like wavelets, reflecting the fast transients in these sounds; whereas the filters for speech were composed of both representations (Figure 7C). By comparing them with neurophysiological data of auditory nerves, it provided an explanation for the particular distribution of cochlear tuning properties (the form of their filtering properties, and their organization as a population) (Smith & Lewicki, 2006)(Figure 7D). Therefore, speech sounds are “special” (but exactly like other natural sounds) because their acoustic properties make efficient use of the bandwidth available in the auditory system, matching its encoding capabilities. The auditory neural code appears to be particularly well-matched to the statistical properties of the speech signal. Auditory neurons could also be tuned so that the acoustic differences between natural sounds are maximized in the brain, thus facilitating auditory discrimination. Whereas, in natural visual scenes, a prevalent form of structure is an edge which can be efficiently encoded by a population of Gabor filters, in the auditory domain, their equivalent are sound onsets or ‘acoustic edges’ that can be efficiently encoded by a population of filters that resembles a gammatone filter bank. The efficient coding hypothesis postulates that these filters are not ‘edge detectors’, but rather that the code is optimized for a more general class of patterns: those with edges and those that vary smoothly (Lewicki, 2002).

Recently, Albouy and colleagues have shown, using functional magnetic resonance imaging, that the extent to which music and speech processing rely on a complementary specialization of two parallel neural systems, one in each hemisphere, could be explained by the fact that each of these domains exploits a distinct sensitivity to spectrotemporal modulation that maximizes the efficiency of encoding of their respective acoustical features (Albouy et al., 2020). This result is in line with the domain-general approach and emphasizes to some extent the potentially shared neural substrates for sounds, speech, and music.

ii. Sensory, motor and sensory-motor account of speech perception

Several theories have been proposed to account for the processes involved in decoding the speech signal. One of these frameworks can be labeled as ***auditory theories of speech perception*** which postulate that speech perception is based solely on auditory processes. It is dependent on the same general mechanisms (not specific to speech) as those underlying hearing and perceptual learning, and are the result of a spectro-temporal analysis of the acoustic signal (Diehl, Lotto, & Holt, 2004; Massaro, 1974; Ohala, 1996). In other words, phonetic invariants are thought to be retrieved and accessed directly from the acoustic signal, which is then processed only by the auditory system. In this perspective, speech perception is thus considered as the recognition of complex acoustic patterns mainly involving the auditory system (Diehl et al., 2004). These auditory approaches are mainly based on the fact that the phenomenon of categorical perception is neither specific to speech nor to humans (Pisoni, 1977; Stevens & Klatt, 1974). Thus, speech perception would be based on general mechanisms common to those involved in auditory processing.

In contrast to auditory theories, initially, ***the motor theory of speech perception*** assumed that information from the speech signal is mapped onto motor representations of the articulatory gestures that give rise to the perceived signal. Thus, there would be a link between perception and speech production. Everyone would have individualized and internal motor representations, which would be recruited for speech perception. Faced with the invariance between the acoustic signal and the underlying phonetic representations, Liberman proposed to consider the invariants not in acoustic terms but in motor terms. We do not perceive the speech sounds themselves but only the articulatory gestures that generate them (Liberman et al., 1967). Since the listener is also a speaker, Liberman hypothesized that she can use her own experience of speech production to recognize the articulatory movements behind the received acoustic signal. The resulting representations are then invariant. According to Liberman and colleagues, this process is accomplished by a specialized module in the listener's motor system which simulates the speaker's articulatory movements and compares the predicted result with the actual perceived signal (analysis by synthesis) (Kuhl, Ramírez, Bosseler, Lin, & Imada, 2014; Liberman et al., 1967). Later, they proposed that the object of perception is not the articulatory gesture itself, but rather the neuromotor control that causes the articulatory gesture (Liberman & Mattingly, 1985). Finally, in 1986, a student of Liberman, Carol Fowler, proposed a less constrained version of Motor Theory, termed ***direct-realist*** (Fowler, 1986). Although also assuming that the object of perception is the articulatory movement and not the acoustic signal, this explanation does not assume access to a specialized module of the motor system. In place of the analysis-by-synthesis process proposed by Liberman, Fowler proposed that speech perception relies instead on a very general mechanism of perceptual learning of statistical regularities in the speech signal.

Current theories tend to favor the existence of a close interaction between the motor and auditory systems in speech perception, defining verbal units not as purely acoustic or

purely motor, but as sensory-motor (Hickok & Poeppel, 2007; Schwartz et al., 2012). ***The Perception for Action Control Theory*** is one of the models arguing for the sensorimotor character of speech units (Schwartz et al., 2012). It is primarily based on the existence of motor and auditory procedural knowledge established during speech acquisition and shared by the speaker and listener. Language acquisition enables the co-structuring of sensory and motor representations (sensory-motor learning by imitation) which results in the development of sensory-motor maps of speech. Production is structured by perception: articulatory gestures are shaped by their auditory and visual perceptual properties. The perceptual system provides the production system with auditory templates containing auditory and visual perceptual properties to realize pertinent articulatory gestures. Finally, the perceptual system is shaped by motor constraints of the orofacial gestures. Thus, according to the Perception for Action Control Theory model, during the perception of a speech signal, the articulatory representations of the listener constrain the processing and interpretation of the sensory inputs through internal motor simulation mechanisms.

2.2 (Neuro)biological perspectives

i The Dual route model

In line with studies showing the involvement of motor regions in speech perception, recent neurobiological models postulate a close link between the speech perception and production systems. One such a model is based on the assumption that to be able to understand and speak any language, there are a least two requirements: a specific architecture dedicated to the transformation of acoustic information into a representation of the meaning of an utterance (auditory-conceptual route), and on the other hand, there is a need for an auditory-motor route, such that speakers can reproduce speech specific sound patterns with their vocal tract. Hickok and Poeppel have proposed that these two processing streams involve partially segregated brain circuits within their ***dual route model of speech processing*** framework (Hickok & Poeppel, 2000, 2007; Poeppel & Hickok, 2004). Their model of speech processing, largely expands on historical works of speech perception and production such as the Classic (outdated) Model, often referred to as the ‘‘Broca-Wernicke-Lichtheim-Geschwind model” (Tremblay & Dick, 2016). However, their view is that language processing is a highly distributed and dynamic process. As the organisation of speech processing is task-dependent, they proposed to distinguish speech perception, which refers to sub-lexical tasks, such as syllable discrimination, from speech recognition (auditory comprehension), which refers to the processing step required in the transformation of an acoustic signal into lexical representations. The first steps in speech processing are realized bilaterally: (i) a spectro-temporal analysis is performed in the bilateral auditory cortices, in the dorsal part of the superior temporal gyrus (STG), with possible

inter-hemispheric computational differences. (ii) the processing of phonetic and phonological representations involves the posterior part of the medial superior temporal sulcus (STS) and is possibly left lateralized (Turkeltaub & Coslett, 2010). The resulting information is then processed in different regions, depending on the cognitive demand.

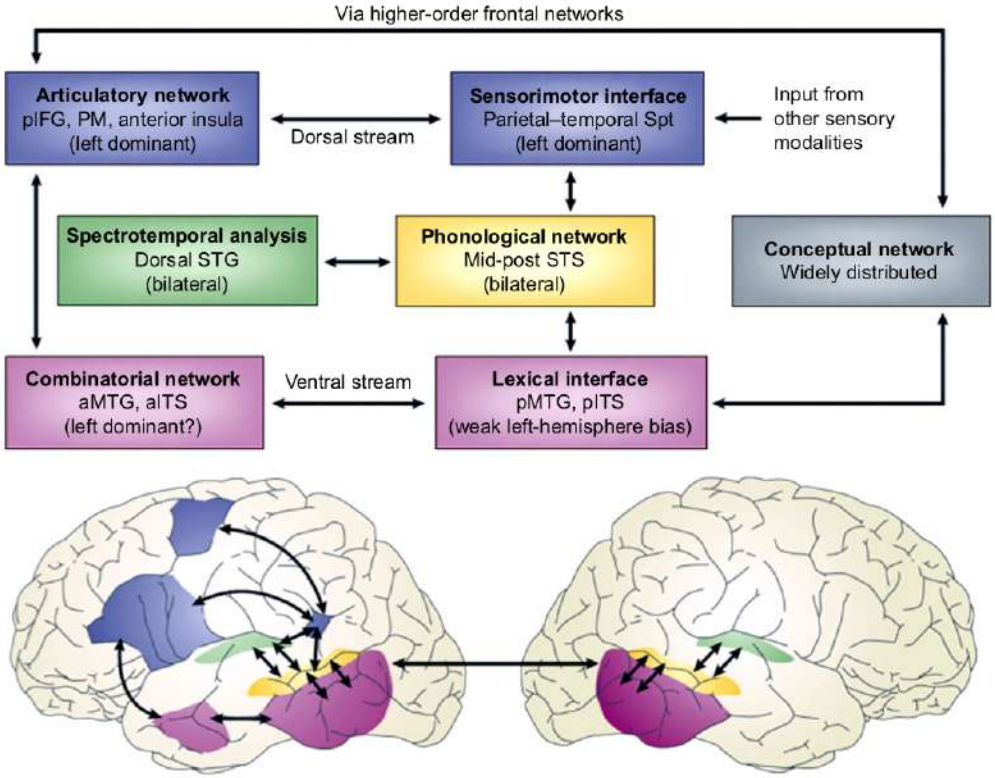


Figure 8. Top panel: The earliest stage of cortical speech processing is carried out in auditory cortices bilaterally in the supratemporal plane. Subsequently, the system diverges into two broad streams, a dorsal pathway (blue) that maps sensory or phonological representations onto articulatory motor representations, and a ventral pathway (pink) that maps sensory or phonological representations onto lexical conceptual representations. Bottom panel: Approximate anatomical locations of the dual-stream model components. Regions shaded green are proposed to be involved in spectrotemporal analysis. Regions shaded yellow are implicated in phonological-level processes. Regions shaded pink represent the ventral stream, which is bilaterally organized with a weak left-hemisphere bias. The more posterior regions of the ventral stream correspond to the lexical interface, which links phonological and semantic information, whereas the more anterior locations correspond to the proposed combinatorial network. Regions shaded blue represent the dorsal stream, which is strongly left dominant. The posterior region of the dorsal stream corresponds to an area in the Sylvian fissure at the parietotemporal boundary (area Spt), which is proposed to be a sensorimotor interface, whereas the more anterior locations in the frontal lobe, probably involving Broca's region and a more dorsal premotor site, correspond to portions of the articulatory network. IFG: inferior frontal gyrus; ITS: inferior temporal sulcus; MTG: middle temporal gyrus; PM: premotor; Spt: Sylvian parietal-temporal; STG: superior temporal gyrus; STS: superior temporal sulcus. Adapted from (Hickok & Poeppel, 2007).

the ventral stream

The ventral stream for speech recognition of acoustic objects contains several routes for lexical access, distributed bilaterally, with a possible left dominance for phonemic labelling, in contrast with non-categorical representations. It should be noted that the different lexical

analysis performed in parallel can be characterised by different computational processes, with a possible complementarity or redundancy. An important question is, on the one hand, to know at which stage of cortical processing these parallel and complementary processes start, and on the other hand, to characterize the processes showing an asymmetry. In this respect an analogy can be made with the redundancy in the acoustic signal of temporal and spectral cues important for comprehension, a sign of robust encoding (Rosen, 1992). Finally, these regions do not specifically process the speech signal at the expense of other acoustic information. The lexical interface between phonological and semantic (distributed) information is located in the posterior portion of the temporal lobes, while the left anterior part corresponds to the combinatorial network, involved in high-level syntactic processing and compositional semantic operations.

the dorsal stream

The dorsal, left-lateralized (but see Cogan et al., 2014) stream allows for auditory-motor integration and underlies phonological short-term memory (Jacquemot & Scott, 2006). It is essential during development to adjust the sounds produced by the child to the patterns heard from others (Doupe & Kuhl, 1999). It contains an anterior, frontal part, probably including Broca's area and dorsal premotor regions, which corresponds to the articulatory network. The posterior part consists of the Spt region, located in the posterior Sylvian fissure at the parieto-temporal junction. It hosts the sensory-motor interface, but seems to be closer to the motor than to the sensory system. Its activity is indeed not specific to speech sounds or auditory materials but seems to be related to the efferent system of the vocal tract. Activity in Spt is highly correlated with activity in the pars opercularis (Buchsbaum, 2001; Buchsbaum, Olsen, Koch, & Berman, 2005), which is the posterior sector of Broca’s region. White matter tracts identified via diffusion tensor imaging suggest that Spt and the pars opercularis are densely connected anatomically (for review see Friederici, 2009; Rogalsky & Hickok, 2011). Moreover, the computational properties underlying auditory-motor integration in the dorsal stream have recently been detailed through the notion of forward models in speech motor control. The basic idea is that to produce speech, the brain constantly predicts the consequences associated with its own movement with the use of an internal model. At each moment in time the brain receives copies of motor commands and integrates them with information about the ongoing state of the motor system. This specific mechanism enables the detection and also correction of the speech production in case of mismatch between predicted motor command and the one actually realized. In the case of speech sounds, inputs to the auditory phonological network define the auditory targets of speech acts to perform. When speaking then, predicted auditory consequences of the speech production can be compared to the auditory target from the phonological network.

The above-mentioned model was first introduced nearly twenty years ago and has since been refined and updated on the basis of experimental data coming from neurophysiological recording in healthy subjects and also patients. Although specific

modifications have been made to account for experimental observations, they do not challenge the core of the model but rather clarify it.

After having introduced several important theoretical as well as more biologically grounded frameworks, let's dive more into the details of the neural architecture and functioning supporting sounds processing including speech.

ii. Processing within the auditory pathway

Sounds from the environment enter the ear and cause the eardrum to vibrate. These vibrations are then transmitted through the middle ear via the three ossicles to the cochlea via the oval window. As these sound waveforms enter the ear, the cochlea decomposes them into time-frequency representations.

The cochlea is composed of ~3,500 inner hair cells and ~12,000 outer hair cells inserted inside the basilar membrane. The movements of the basilar membrane results in the firing of the inner hair cells. Two different codes are used by the cochlea to transduce the acoustic signal: a "spatial code" and a "temporal code".

- ***Spatial code.*** The entering sound waves will deform the basilar membrane at a frequency-dependent location. High frequencies (20 kHz) act at the base of the cochlea and low frequencies (20 Hz) at the apex. This is called cochlear tonotopy. The inner hair cells are inserted along the cochlea and thus also respects a tonotopy. Cells located at the basis of the membrane fire when low frequency sounds are presented while cells located at the apex of the membrane fire for high frequency sounds.
- ***Temporal code***. The inner hair cells can also synchronize with the frequency of the acoustic stimulation. For example, cells fire at 100 Hz when presented with a 100 Hz pure tone. But due to physiological properties of the hair cells, this code is limited to ~ 5kHz.

Following processing in the cochlea, the tonotopic information is preserved within the auditory nerve and throughout the entire ascending auditory pathway into the core auditory cortical regions.

Hearing, like any other sensory modality, has a primary pathway and centres which are entirely dedicated to this function. This pathway is rather fast with large myelinated fibres and leads to the primary auditory cortex. The first relay of the primary auditory pathway is the cochlear nucleus of the brainstem which receives the axons of the type I neurons of the spiral ganglion (auditory nerve); at this level an important processing is done in the basic decoding of the message: duration, intensity, frequency.

A second major relay in the brainstem is the superior olivary complex: most auditory fibres make a synapse here after having crossed the median line. From this relay, a third neuron sends the message up to the inferior colliculus in the midbrain. These two relays play an essential role for sound localisation. At this stage, neural responses are also more robust to noise and reverberations thanks to dedicated processes. A final relay, before the cortex, is

carried out in the medial geniculate body in the thalamus; Here, the upcoming information is integrated with other sensory information such as visual, somatosensory and vestibular. The last neuron of the primary auditory pathway links the thalamus to the auditory cortex where the message, already largely decoded by upstream neurons, can be further analysed and integrated. Overall, auditory neurons present in subcortical nuclei, including the cochlear nucleus and inferior colliculus, phase-lock very well with stimuli presented at rates over 200 Hz (Nourski & Brugge, 2011). However, the capacity for envelope-locking decreases along the auditory pathway as demonstrated by recording in the medial geniculate body (16 Hz), the primary auditory cortex (8 Hz) and non-primary auditory areas (4–8 Hz) (Arnal, Poeppel, & Giraud, 2016).

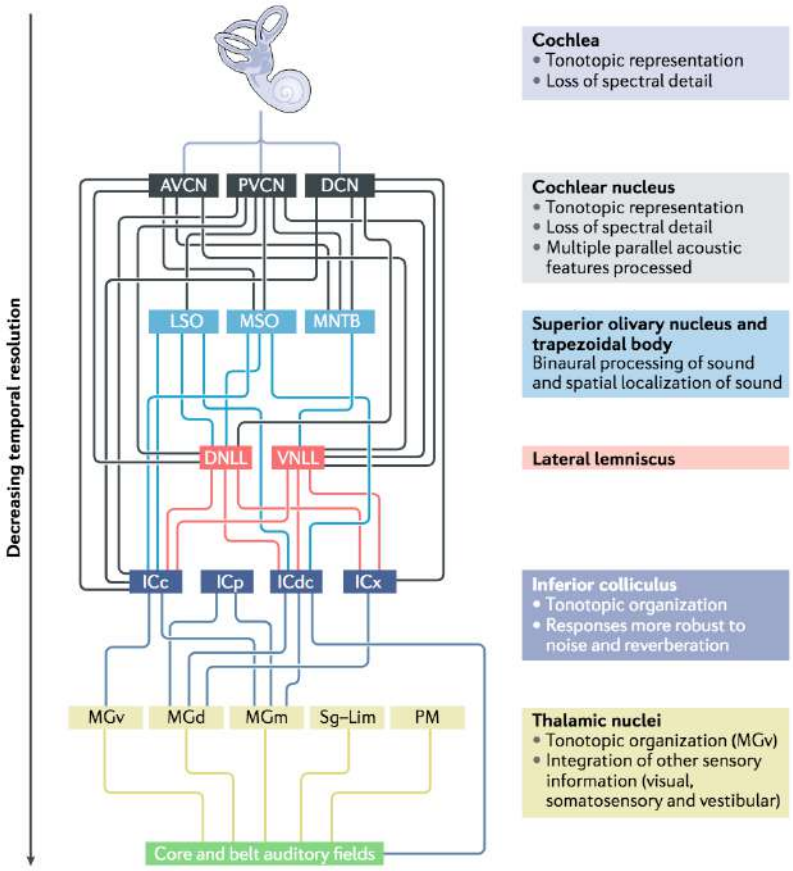


Figure 9. Schematic representation of the ascending and descending auditory pathways with all relays and their putative role in auditory processing. Starting from the cochlea (top) to core and belt auditory regions (bottom). Adapted from (Jasmin, Lima, & Scott, 2019).

There exists also a non-primary auditory pathway. After the first relay in the cochlear nuclei, small fibres connect with the reticular formation where the auditory message joins all other sensory messages. After several relays in the reticular formation, then in the non-specific thalamus, this pathway leads to the associative cortex. The main role of this pathway, also connected to wake and motivation centers as well as vegetative and hormonal systems, is to select the type of sensory message to be processed first. Of note, the 'non-specific' thalamic afferents are widespread and project both within and across sensory modalities. Such pathways have notably modulatory effects on sensory processing which can be implemented in the reorganisation of ongoing neuronal oscillations in supragranular layers of the targeted cortical regions (Schroeder & Lakatos, 2009).

In the auditory system, similar to many sensory and motor pathways, the primary route from the periphery to the brain is crossed (Moore, 1987). Although there exist ipsilateral pathways between the periphery and the brain, sounds entering the right ear are mostly processed in the left hemisphere and vice versa. This is due to the strong crossed pathways between the ear and the auditory cortex.

Modelling the multiple processing steps occurring subcortically has been instrumental in studying sound and speech perception. Current models include a gammatone filterbank that models the cochlea and its mechanotransduction (Lyon & Katsiamis, 2010; Patterson & Nimmo-Smith, 1987). The output of this stage is then nonlinearly processed using half-wave rectification or Hilbert transform (Ainsworth et al., 2012; Hudspeth, 2008). The resulting signal is further passed to a modulation filterbank and a detector mechanism (Xiang, Poeppel, & Simon, 2013). Whereas we currently have pretty good biologically-inspired models of the first processing steps along the auditory hierarchy, modelling the processes at play within the auditory cortices remains a major challenge.

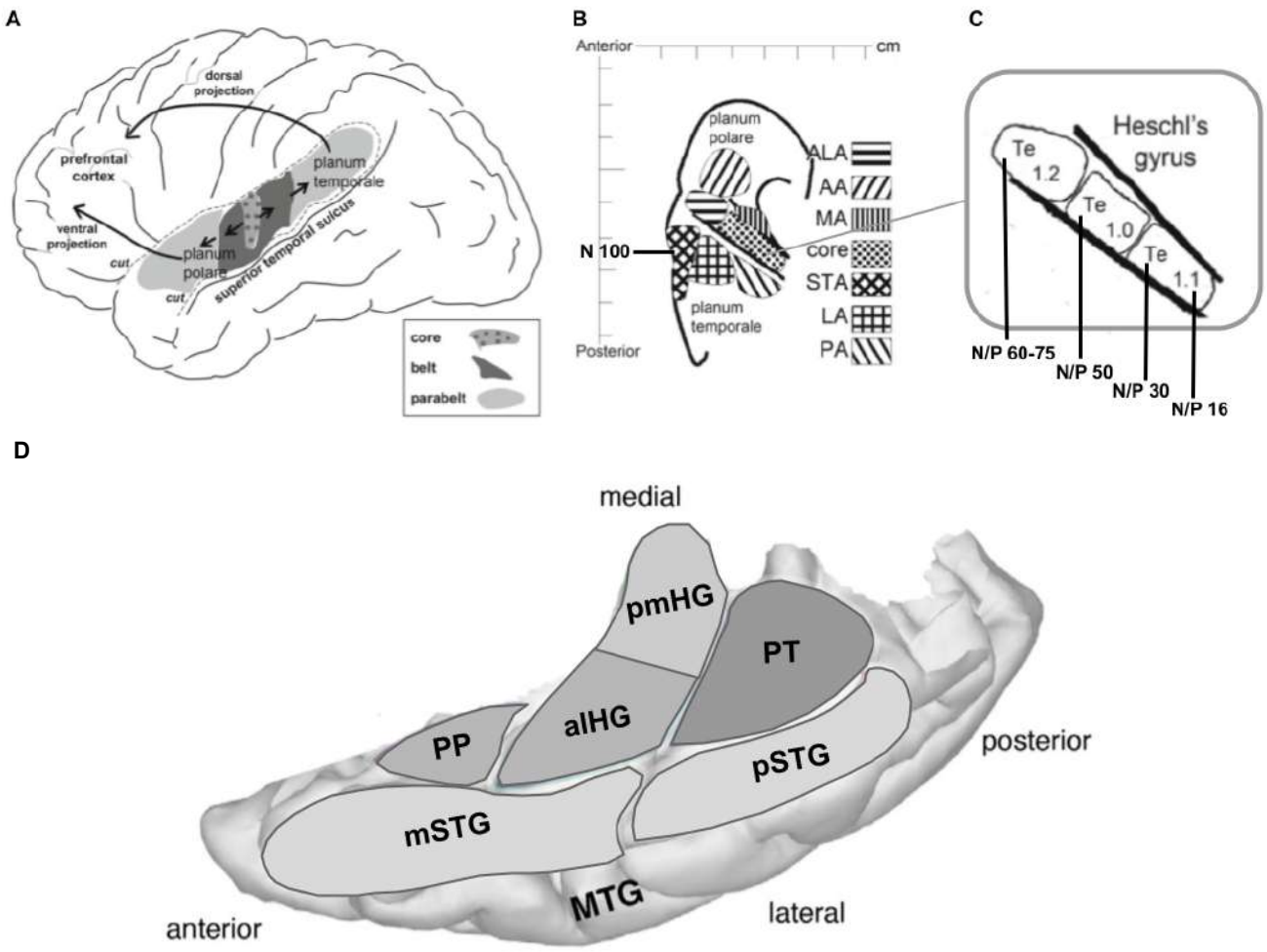


Figure 10. A) Surface of human left hemisphere with a cut through the Sylvian fissure to reveal the macroanatomical structure of auditory cortex on the inner surface, including Heschl's gyrus, planum polare and planum temporale. In this panel, the position of Heschl's gyrus (the core region) is shown by the dotted grey region. A suggestion for how belt and parabelt regions might be organized is shown by the dark (belt) and light grey (parabelt) shading. B) Summary diagram of the microanatomical structure of the human supratemporal plane (left hemisphere). C) Three areas Te1.1, Te1.0, and Te1.2, within the primary auditory cortex (Brodmann area 41) can be identified along the mediolateral axis of Heschl's gyrus. Anatomical localization (schematic) of the generators of the medium latency auditory evoked potential components along Heschl's gyrus recorded using intracranial electroencephalography. Letter indicates polarity (N: negative; P: positive), number indicates latency (in milliseconds). D) Anatomical parcellations of temporal lobe regions of the human auditory cortex within the left hemisphere. **ALA**: anterior lateral area, **AA**: anterior area, **MA**: medial area, **core**: Heschl's gyrus, **STA**: superior temporal area, **LA**: lateral area, **PA**: posterior area, **PP**: Planum Polare, **PT**: Planum Temporale, **STG**: superior temporal gyrus, **MTG**: middle temporal gyrus. **pmHG**: posteromedial Heschl's gyrus, **alHG**: anterolateral Heschl's gyrus, **pSTG**: posterior superior temporal gyrus, **mSTG**: middle superior temporal gyrus. A) adapted from (Poeppel, Overath, Popper, & Fay, 2012); B) adapted from (Wallace, Johnston, & Palmer, 2002); C) adapted from (Morosan et al., 2001) and (Liégeois-Chauvel, Musolino, Badier, Marquis, & Chauvel, 1994); D) adapted from (Hamilton, Oganian, Hall, & Chang, 2021).

The auditory cortical system lies in one of the most folded regions of the brain, with great variability between individuals but also between hemispheres. As a result, and due to its particular localization and organization, it has been more challenging to characterize at a

fine-grained scale the auditory cortex than the visual cortex for example. Neuropsychology as well as post-mortem studies have been instrumental in gaining insights into the functioning of the auditory cortex. Currently, most investigations in humans make use of non-invasive recording techniques such as functional magnetic resonance imaging or electro- or magneto-encephalography. Recently, more invasive methods have been also used and include intracranial electro-encephalography and electrocorticography. These specific recording schemes are only used in particular clinical conditions such as refractory epilepsy. They allow a more spatially and temporally characterization of the neural activity but are restrained by clinical considerations. In parallel, invasive recordings have been performed in animals. In mice, finches or monkeys, it is thus possible to have access to unitary activity. Thanks to such a repertoire of investigative tools, we are beginning to have a better understanding of the cortical fields involved in hearing.

Despite the fact that the human auditory cortex has been studied for more than a century, currently, there is no consensus about its anatomical organization. Brodman described the primary auditory cortex as a koniocortical area (41) in Heschl's gyrus. Since then, in this region, around twenty different cytoarchitectonic areas have been distinguished, reflecting the uncertainty surrounding the exact boundaries of these auditory areas (Wallace et al., 2002). It is now agreed that the primary auditory cortex is located within the lateral fissure, on the supratemporal plane of the transverse temporal (Heschl's) gyrus. The primary auditory cortex has been located in the medial two-thirds (Liegeois-Chauvel, Musolino, & Chauvel, 1991), in the posteromedial two-thirds (Wallace et al., 2002) or in the mediolateral axis of Heschl's gyrus (Morosan et al., 2001). Heschl's gyrus exhibits a highly variable morphology which may include one to three gyri per hemisphere, with the number of gyri potentially different between hemispheres (Marie, Maingault, Crivello, Mazoyer, & Tzourio-Mazoyer, 2016).

The primary auditory cortex is the central region of the auditory cortex and receives direct projections from the ascending auditory pathway, especially from the ventral region of the medial geniculate body in the thalamus. The primary auditory cortex is composed of neurons involved in decoding the cochleotopic and tonotopic spatial representation of a stimulus. It contains neurons selective for particular frequencies and arranged in isofrequency bands that are tonotopically organised. Their activity is stimulus-dependent: and as such, varies as a function of the frequency, intensity and position of the sound source in space. Thanks to the level of granularity of the methodology, recent intracranial work has described spatial gradients for frequency, response latency, temporal modulation and phonemic encoding in the Heschl's gyrus of epileptic patients (Khalighinejad et al., 2021). Moreover, in the primary auditory cortex, the firing rate of a large number of neurons is phase-locked to slow temporal modulations below 30 Hz (Liang, Lu, & Wang, 2002; Malone, Scott, & Semple, 2010; Yin, Johnson, O'Connor, & Sutter, 2011). Although single auditory cortical neurons generally cannot phase lock to temporal modulations above 100 Hz (Wang, 2007), macro-scale measurements such as local field potentials and human magneto-encephalography show phase locking to the stimulus modulations beyond 100 Hz,

reflecting the ability of auditory cortical regions to maintain a representation of higher temporal frequencies (Pesnot Lerousseau, Trebuchon, Morillon, & Schon, 2019).

The precise number of non-primary auditory cortices remains unknown at present. Brodman determined that the primary auditory cortex was surrounded caudolaterally by area 42, rostrolaterally by area 22 and medially by area 52; representing three secondary auditory areas. More recently, studies have identified one or two core areas surrounded by at least six belt areas (Rivier & Clarke, 1997; Wallace et al., 2002). Based on gross morphological criteria, the non-primary human auditory areas have been partitioned into the planum polare, corresponding to the area anterior to the Heschl's gyrus, and the planum temporale, which is posterior to the Heschl's gyrus (Hall, Hart, & Johnsrude, 2003). In comparison to the primary auditory cortex, the non-primary auditory fields do not have a clear tonotopic organisation but have an important role in sound localisation, integrating hearing with other sensory systems and in the analysis of complex sounds such as human speech.

Most complex sounds contain crucial spectral and temporal information. Accumulating evidence suggests that the human auditory cortex is tuned to spectro temporal modulation rate, as shown by the existence of spectrotemporal receptive fields (Chi et al., 2005). Schönwiesner & Zatorre, using fMRI and dynamic ripple stimuli, characterized a representation of spectrotemporal modulations in the human auditory cortex (Schönwiesner & Zatorre, 2009). They showed that within both primary and non-primary auditory cortices, some voxels were highly selective to spectral and temporal modulations. This work implies that a combination of temporal and spectral modulation is highly relevant for auditory processing. Another study by Santoro and collaborators corroborates these findings by showing that the representation of natural sounds in the human auditory cortex relies on a frequency-specific analysis of combined spectro-temporal modulation (Santoro et al., 2014). This specific representation of sounds at multiple resolutions may be the key requirement to form auditory objects and comprehend complex signals such as speech (Formisano, De Martino, Bonte, & Goebel, 2008). The use of more ecological speech stimuli (audiobooks, discourses) coupled to neuroimaging recordings has enabled the characterisation of the spatio-temporal profile of brain activity during spoken language comprehension (Donhauser & Baillet, 2020; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016). These recent studies support the view that speech comprehension involves hierarchical representations that start in primary auditory areas and move laterally along the temporal lobe.

Different types of experimental material have been used to gain insight into auditory processing in humans. Ranging from highly artificial and controlled pure tones, to various but precisely defined types of noises to more naturalistic stimuli such as syllables or even long excerpts of continuous speech. Such a variety of stimuli has fruitfully extended our comprehension of diverse and complementary mechanisms at play during auditory perception. The analysis of speech, especially at the first processing steps have been shown to rely on general and shared auditory mechanisms that are also relevant for processing any other kind of sounds (Moore, 2008).

The development of neuroimaging techniques has also largely contributed to a better characterization of the cortical auditory territories. Human superior temporal gyrus has been described as a critical cortical structure that sits at the interface between lower-level auditory and higher-level association areas that support abstract aspects of language. It is assumed that this territory supports the encoding of more complex and heterogeneous sound features compared with earlier regions in the auditory hierarchy. It has been shown that its activity reflects the local encoding of acoustic-phonetic features (Mesgarani et al., 2014). More generally, it is likely that these results reflect the region specific sensitivity to complex spectrotemporal tuning. There is also evidence for a non-linear encoding of acoustic input as demonstrated by particular response profiles corresponding to categorical phoneme percepts (Chang et al., 2010). Recently, the spatial organization of superior temporal response has been characterized using speech stimuli. Its posterior portion is sensitive to speech onset following a period of silence, whereas the middle-to-anterior part shows tracking of ongoing changes in the amplitude envelope of continuous sound (Hamilton, Edwards, & Chang, 2018). Onset responses have been found for both intelligible and unintelligible speech as well as for non-speech sounds, suggesting that they may reflect a general mechanism of auditory computation (Hamilton et al., 2018). As onsets of speech sounds are characterized by rapid increases in amplitude and as the posterior superior temporal gyrus responds preferably to such cues, it is believed to be involved in the processing of high temporal modulations (Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016). In contrast, its middle-to-anterior portion has been preferably implicated in high spectral modulation processing with characteristic sustain responses (Hullett et al., 2016). The superior temporal gyrus has a critical role for interpreting auditory input and generating higher-level perceptual representations. Indeed, evidence for various forms of contextual influence on its activity is accumulating. Recent studies reported a rapid and dynamic modulation of the superior temporal gyrus neural populations activity following variation in preceding perceptual experiences. These effects potentially reflect influence exerted by the predictability of longer-timescale phonological, lexical, and semantic knowledge (Blank, Spangenberg, & Davis, 2018; Holdgraf et al., 2016; Khoshkhoo, Leonard, Mesgarani, & Chang, 2018). Furthermore, whereas neural populations in the human primary auditory cortex are not sensitive to the intelligibility of speech as reported by listeners, neural populations throughout lateral superior temporal gyrus have been shown to respond stronger to intelligible sounds, demonstrating their implication in the perceptual representation rather than only sensory experience (Nourski et al., 2019). Chang argues for recurrent processing mechanisms as being at the core of the ability of superior temporal gyrus to incorporate and bind sensory information at multiple timescales to efficiently represent linguistic information at multiple levels of representation (Chang et al., 2010). This specific computation may be at the basis of the still mysterious mechanism enabling the shift from an acoustic to a linguistic analysis of the speech signal.

Recently and based on intracranial recordings, electrocortical stimulation, and surgical ablation, Hamilton and colleagues demonstrated that cortical processing across auditory

areas was better characterized by a parallel and area-dependent organization than consistent with the serial hierarchical dominant view (Hamilton et al., 2021). By simultaneously recording neural activity from multiple subfields of the auditory cortex using electrocorticography, they found evidence for distributed and parallel processing, with early latency responses observed throughout the posterior temporal plane and superior temporal gyrus. Moreover, thanks to direct cortical stimulation and an ablation case study they provided evidence that Heschl's gyrus is neither necessary nor sufficient for speech perception in humans. This stance of work complements previous findings in monkeys, where the existence of parallel processing simultaneously with hierarchical processing has been well demonstrated (Rauschecker, 1998). Furthermore it confirms in humans what has been found in animals; that is, after a lesion of the primary auditory area in the monkey, the responses recorded in its adjacent region: the rostrolateral auditory area, remained unchanged suggesting that this area received afferents independently of the primary area. By contrast, the responses recorded in the caudomedial area appeared to be dependent on the integrity of the primary auditory cortex (Rauschecker, 1998).

iii Neural oscillations and speech processing

One of the fundamental properties of brain activity is the presence of rhythmic fluctuations at different frequencies. These brain oscillations result from the periodic synchronous electrical activity of a population of neurons at a certain frequency (Berger, 1929). It has been proposed that populations of neurons synchronise their activity dynamically and that these oscillations reflect cyclical variations in neuronal excitability (alternating between phases of high excitability (hyperpolarization) and low excitability (depolarisation); (Buzsáki & Draguhn, 2004; Lakatos et al., 2005). Oscillations have been described in the brains of many mammals (Buzsáki, Logothetis, & Singer, 2013). Neural oscillations are observed both at rest (they are therefore intrinsic, spontaneous and independent of any external stimulation) and during cognitive, perceptual or sensory-motor activity. This synchronisation indicates that different assemblies of neurons discharge at the same time and are hypothesized to reflect optimal communication (temporal) windows, favoring information processing (Fries, 2015; Varela, Lachaux, Rodriguez, & Martinerie, 2001).

Large-scale synchronisation with external stimuli can also be observed: brain oscillations can align their phase or their amplitude with the phase of a rhythmic, periodic or quasi-periodic external stimulus (such as speech or music) (Gross et al., 2013; Pesnot Lerousseau, Trébuchon, Morillon, & Schön, 2021). This type of synchronisation is referred to as 'entrainment' (Lakatos et al., 2005; Schroeder & Lakatos, 2009). Upon the arrival of an external stimulus with a (quasi-)rhythmic structure, a neural response is generated in the regions involved in processing the stimulus. The quality of this response, and therefore the quality of the processing of the information, was observed to depend on the oscillatory state of the neuronal system at the time the information is presented (Schroeder & Lakatos, 2009). If the information arrives in the phase of high excitability of the neuronal ensemble, the rate of

neuronal discharge will be greater and the processing will therefore be optimal. Conversely, if the stimulus arrives in the opposite phase (at low excitability), the rate of discharge will be lower and may even suppress communication within the neuronal networks involved, thereby decreasing the efficiency of information processing (Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007; Lakatos et al., 2005). Neuronal oscillations can also reset or phase shift to the phase of the periodic or quasi-periodic stimulus (Lakatos et al., 2007; Lakatos et al., 2005; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). This "entrainment" would amplify the neuronal response and thus "boost" the effectiveness of the processing. In other words, if information from the environment has a (quasi-)regular temporal structure, it can be processed optimally (e.g., more quickly and efficiently) if the neuronal oscillations in the region(s) involved in processing align so that the occurrence of sensory information coincides with peaks in neuronal excitability (Peelle & Davis, 2012; Schroeder & Lakatos, 2009). From this perspective, the alignment between neuronal oscillations and (quasi-)rhythmic information would therefore be a fundamental mechanism for ensuring optimal sensitivity of our organism to perceptual events (Peelle & Davis, 2012).

Schroeder and Lakatos have proposed that the brain makes use of different processing strategies depending on the specific characteristics of the incoming sensory information. Based on the observation that in natural settings, stimuli often occur in rhythmic streams (our environment being dynamics and presenting temporal regularities, speech, music and motor movements being the most striking examples), they postulated that in those cases, brain oscillations can entrain to the input rhythm and align their high excitability phases with the events in the stream, resulting in an amplification of the neuronal responses and suppression of responses to events that occur out of phase (e.g., that are not part of the attended stream). In this 'rhythmic mode' neuronal oscillations within primary sensory cortices are key elements of the active sensory input selection. In contrast, when no low-frequency environmental rhythms are present, and thus that neuronal entrainment cannot happen, attention operates in a 'continuous mode' which is supported by the suppression of low-frequency neuronal oscillations and a state of high excitability characterized by an extended increase in gamma synchrony (Schroeder et al., 2008).

Noteworthy, endogenous oscillatory dynamics exist in auditory cortical regions, at rest, potentially revealing auditory processing constraints (Giraud et al., 2007; Keitel & Gross, 2016; Lubinus et al., 2021; Morillon, Liégeois-Chauvel, Arnal, Bénar, & Giraud, 2012). Guided by these observations but also by previous line of research (Suppes, Lu, & Han, 1997), Luo and Poeppel showed that a set of spoken sentences could be distinguished from each other thanks to the time series of the phase of the low-frequency brain activity from human auditory areas. They concluded that dynamic brain activity in the theta band (4–8 Hz) temporally aligns or becomes phase resets to the dynamic regularities in speech (Luo & Poeppel, 2007). Since then, an incredible amount of scientific work has been carried out in order to explicit and precisely define the underlying neural mechanisms of speech entrainment or tracking and their implications for speech comprehension. Results from a cocktail-party paradigm (multiple talkers situation) reveal that when listeners are specifically attending to only one speaker in

such a situation, whereas the neural tracking of acoustic features reflects a mixture of both talkers, at the lexical level of representation, the neural tracking reflects exclusively the attended speaker (Zion Golumbic et al., 2013). Such data are in accordance with the idea that attentive listening can selectively affect specific encoding of acoustic or linguistic representations (Brodbeck, Hong, & Simon, 2018). By carefully orthogonalizing acoustic and linguistic timescales within artificial speech stimuli, Ding and colleagues were able to demonstrate that the brain activity reflected the tracking of linguistic structure which was not present acoustically (Ding, Melloni, Zhang, Tian, & Poeppel, 2016). These results speak in favor of neural tracking of speech as a mechanism or at least a relevant index of speech perception.

iv Processing constraints

A potential way of uncovering general processing principles useful in the case of speech perception might be to study and determine the constraints limiting the speech processing system. A large body of work has sought to characterize such limiting factors and result from more general observations of the characteristics of auditory processes in diverse sound environments.

Multiple everyday life examples illustrate the ability of our auditory system to adapt to degraded listening situations. For instance, imagine being in a noisy pub abroad listening to a friend who is not a native speaker of your language. In this situation, your perceptual system is confronted with several challenges: non pertinent source of external noise, specific accent that modify the relationship between acoustic cues and linguistic representation and perhaps also fast speech rate. However, you are (mostly) able to understand what your friend is saying. The mechanisms behind this astonishing feat have started to be investigated in detail.

At the behavioral level, several studies have shown that humans can understand spoken sentences presented at up to three times the normal speech rate, demonstrating the robustness and the perceptual flexibility of the language comprehension system to acoustic variations (Ahissar et al., 2001; Dupoux & Green, 1997). However, at higher rates, comprehension drops abruptly (Peelle, McMillan, Moore, Grossman, & Wingfield, 2004). Previous research using speeded speech has provided evidence that beyond 15 syllables per seconds, speech becomes unintelligible. Noteworthy, exposure to time-compressed speech also leads to a perceptual after-effect: normal-rate speech sounds unnaturally slow immediately after listening to time-compressed speech (Davis et al., 2018). These classical behavioral observations imply that the neural architecture underlying speech processing operates within a range compatible with the typical speech rate but is also able to flexibly adapt to a certain degree to statistical properties of the speech signal in more extreme conditions.

Neuroimaging studies have shed light on the neural substrates supporting this ability. Studies found that in specific brain regions, such as Heschl's gyrus and the neighboring territories of the superior temporal gyrus, neural activity tracks the temporal envelope of compressed speech, even when comprehension breaks down (Nourski et al., 2009;

Vagharchakian, Dehaene-Lambertz, Pallier, & Dehaene, 2012). In contrast, other brain areas including the anterior part of the superior temporal sulcus showed a consistent response while speech remained unintelligible but ceased to respond for compression levels that were no longer understandable (Vagharchakian et al., 2012). The diversity of response patterns can be explained by the existence of different temporal receptive windows across brain regions. With shorter ones in primary and sensory regions and longer ones in higher-level regions such as anterior and posterior superior temporal gyrus and inferior frontal gyrus. Such a mechanism would allow the integration or chunking of the input along progressively longer temporal scales at the same time providing the neural systems with hierarchical memory buffers. Several studies, using neurophysiological recordings, have demonstrated the role of the theta rhythm (4-8 Hz) for adaptation to speech rate. When listening to time-compressed speech, oscillatory activity is phase locked to the envelope of the sound in the theta band (Ahissar et al., 2001; Pefkou, Arnal, Fontolan, & Giraud, 2017). Phase-locking to the envelope of sentences was also observed in the high-gamma band during intelligible but also unintelligible speech (Nourski et al., 2009). These results thus highlight the fact that while envelope-tracking by cortical activity is a prerequisite to adapt to time-compressed speech, it is in itself insufficient to explain comprehension. In a recent magnetoencephalography study Lizarazu and colleagues recorded neural responses from participants while they were listening to speech at decelerated, normal, and accelerated rates (Lizarazu, Lallier, & Molinaro, 2019). They found that the phase of theta band oscillations in the auditory cortex adjusts to speech rates. Moreover, they showed that the coupling between the phase of the theta activity and the frequency peak of the amplitude of the gamma oscillation follows the speech rate. Their results suggest a crucial role of the theta-gamma coupling in the initial encoding process of speech comprehension.
More studies are needed to specify the extent to which neural speech tracking can adapt and to precise its link to speech comprehension. Other mechanisms may also be necessary to provide a full account of time-compressed speech perception (see part 2.4 current models).

2.3 An asymmetrical organisation

i. Spatial & structural asymmetries

Since the middle of the 19th century, consistent experimental evidence has highlighted the remarkably and ubiquitously asymmetrical nature of the brain and, in fact, asymmetry, rather than symmetry, seems to be the norm in brain function. Technological developments in structural and functional imaging have enabled detailed characterization of the brain. Auditory and language areas have been specifically examined and have revealed asymmetries at multiple levels of investigation. For instance, the Planum Temporale has been described as the most asymmetrical cerebral structure in the whole brain (Prete, Marzoli, Brancucci, & Tommasi, 2016). Structural asymmetries can be found at different scales including at the

macrostructural (volume or thickness), microstructural (cell number, cell packing, dendritic arborization) but also at more integrated (networks, fibers) levels.

In 1878, Heschl, in his description of the transverse anterior gyrus, already suggested the existence of asymmetries in cortical folding between the transverse and superior anterior temporal gyrus (Heschl, 1878). It appears that the volume of the left Heschl's gyrus is directly correlated with the extent of cortex involvement in temporal processing of sounds, while its right counterpart's volume correlates with the extent of spectrally related activity in right auditory areas (Warrier et al., 2009). Geschwind and Levitsky work on the planum temporale of 100 adult brains provided the first quantitative measurement of anatomical asymmetry, lateralized to the left in 65% of the brains studied (Geschwind & Levitsky, 1968). These results were confirmed on the brains of fetuses and newborns (Wada, 1969). The asymmetry appears to be greatest, with a ratio of 1 in 7, within the temporoparietal transition region Tpt (Geschwind & Galaburda, 1985), a subpart of which is the key region Spt of the dorsal pathway of Hickok and Poeppel (Hickok & Poeppel, 2007). These results seem to account for the fact that the Sylvian fissure is longer and less angular on the left (LeMay & Culebras, 1972). More recently, a study has shown that the left primary auditory cortex contains more white matter than the right (Penhune, Zatorre, MacDonald, & Evans, 1996). This difference could be due to the greater number of fibres originating from the adjacent planum temporale. These asymmetries have also been extended to grey matter within Heschl's gyrus and the planum temporale (Dorsaint-Pierre et al., 2006), with the ratio of white to grey matter showing a similar inter-hemispheric asymmetry for both regions (Anderson, Southern, & Powers, 1999). Meyer and colleagues investigated the structural asymmetries within five anatomically defined regions of auditory cortex (Heschl's gyrus, Heschl's sulcus, planum temporale, planum polare, superior temporal gyrus) in 104 participants using volumetric 3-dimensional magnetic resonance imaging (Meyer, Liem, Hirsiger, Jäncke, & Hänggi, 2014). They compared bilaterally within the auditory fields, the cortical volume, thickness, and cortical surface area. They found a strong leftward asymmetry for cortical surface area in all auditory regions investigated but planum polare. According to the authors, these findings reflect the existence of a larger number of cortical columns being also more widely spaced apart. They also found a rightward asymmetry in cortical thickness for both Heschl's gyrus and sulcus. Finally, a recent study, focusing on subregions of the planum temporale in a large cohort of more than 350 participants, has shown a positive correlation between an asymmetry of the posterior part of the planum temporale and lexical performances (symmetrical posterior planum temporale being associated with worst performance) (Tzourio-Mazoyer & Mazoyer, 2017). Using in vivo neurite orientation dispersion and density imaging, Ocklenburg and colleagues quantified the axon and dendrite complexity of the left and right planum temporale in 98 participants. Complementary to the previous study, they found that a higher density and higher number of synaptic contacts in the planum temporale increased the temporal precision and reduced the latency of the neurophysiological processing of speech in this area (Ocklenburg et al., 2018).

At a finer level of analysis, there is also evidence for structural asymmetries in cortical auditory fields. The structure of layer III pyramidal cells has been studied in several regions because these cells are thought to be at the origin of long-distance cortico-cortical connections: in primary, secondary and associative auditory areas, as well as in Broca's area. These studies do not show hemispheric differences in the number of cells, but rather a left dominance in the size of large magno pyramidal cells (Hayes & Lewis, 1996; Hutsler, 2003; Hutsler & Gazzaniga, 1996). These large cells therefore are thought to form larger cell columns, connecting adjacent regions with thicker and more branched axons. These results do not seem to extend to the adjacent regions outside the auditory pathway. It has been shown within the primary auditory cortex and planum temporale that the width of a cell column and the spacing between two such columns is greater on the left (Seldon, 1981). These are grouped into macro-columns, and in the planum temporale they are similar in size between the two hemispheres (thus containing fewer micro-columns on the left); however, two macro-columns interconnected via pyramidal cell projections are more widely spaced in the left than in the right hemisphere (Galuske, Schlote, Bratzke, & Singer, 2000). Finally, myelination is also more developed on the left than on the right (Anderson et al., 1999). It should be noted that increased spacing between columns associated with longer dendritic connections could result solely from a dilation of the structure, without any notable architectural and computational modification. However, the left associative cortex has more independent networks than its right counterpart, which are composed of interconnected macro-columns, which goes against this hypothesis. Similarly, adjacent (micro) cell columns are less interconnected in Wernicke's area, leading to less redundant and more elaborate processing. Thus, the connectivity within the micro- and macro-columns (decreased on the left) suggests that the left auditory associative regions have a more refined architecture, with more independent functional units, allowing for a finer analysis of the information.

Investigations measuring structural connectivity have found that the arcuate fasciculus (the white matter fiber bundle connecting Wernicke’s and Broca’s areas which are key regions in the language network) is more developed in the left than the right hemisphere, and this was related to behavioural differences (Catani, Jones, & ffytche, 2005; Catani et al., 2007; Poeppel & Assaneo, 2020). Moreover, a study using graph theory also specifies that globally the right hemisphere has a higher density of connections, which may result in a greater capacity for integration (Sanabria-Diaz et al., 2010). Recently, using graph theory spreading activation models on magnetic resonance imaging data, Mišić and colleagues demonstrated another asymmetry of structural connectivity taking the form of a generally greater integration of the right auditory cortex into the wider connectome than the left, particularly with respect to interhemispheric connectivity (Mišic et al., 2018). The study revealed that this network specificity resulted in differential spreading activation patterns emanating from the left versus right auditory cortex.

In conclusion, these studies indicate that based on differences in architecture, different computational analyses seem to be at play in the left and right auditory areas. However, the

precise relationships between these anatomical peculiarities and the functional differences of both hemispheres are still to be refined.

ii. Functional asymmetries

The role of subcortical auditory pathways in implementing lateralized processing of stimuli is still unclear. Indeed, due to extensive corticofugal influences in the auditory system, the source (at the subcortical or cortical level) of such asymmetries is difficult to discern. However, the peripheral auditory system seems to be characterized by a functional asymmetry in favor of the right ear. For instance, audiometric hearing thresholds in adults are slightly better in the right ear than in the left and this finding is generally limited to males as hearing sensitivity in females is more symmetric (Kannan & Lipscomb, 1974; Pirilä, Jounio-Ervasti, & Surri, 1992). Moreover, the otoacoustic emissions which are a measure of cochlear outer hair cells activity can be measured in the ear canal with a microphone. In general, when they are elicited by transient stimuli such as clicks, the amplitude is larger in right ears than in left (Driscoll, Kei, & McPherson, 2002). Furthermore, at the central level, auditory brainstem responses are greater when the right ear is stimulated than when the left ear is stimulated (Levine & McGaffigan, 1983). Although there is still no clear picture of subcortical functional asymmetry, at the cortical level, the caracterisation is more detailed thanks to diverse experimental data.

One of the first pieces of evidence of cortical functional asymmetry in language processing was brought by the pioneering work of Penfield (Penfield & Boldrey, 1937). He revealed that direct electrical stimulations of the left hemisphere resulted in speech impairments while it was rarely the case upon stimulation of the right hemisphere. In addition, hallucinations and auditory illusions were induced more frequently by stimulation of the right temporal cortex than by stimulation of the left temporal cortex. The ***WADA*** test also led to the characterisation of the differential involvement of the two hemispheres in language processing (Rasmussen & Milner, 1977). This procedure is based on the intracarotid injection of sodium amytal. It results in the temporally inactivation of the hemisphere ipsilateral to the injection. In the language dominant hemisphere, this invasive procedure transiently blocks speech. As such, this test, still in use today in clinical settings, enables the determination of the language dominant hemisphere.

Nowadays, thanks to the development of brain imaging methodologies, many studies have reported functional asymmetries of auditory areas. For instance, previous findings reported that left and right cortical auditory regions exhibit differences in their intrinsic oscillatory activity (Giraud et al., 2007; Morillon et al., 2012, 2010). A relative leftward dominance of low-gamma neural oscillations and/or rightward dominance of theta oscillations is also visible during sensory stimulation (Gross et al., 2013; Lehongre, Morillon, Giraud, & Ramus, 2013; Morillon et al., 2010). This asymmetry is, moreover, reflected in the sensitivity

of the left and right auditory cortex to different spectrotemporal modulations of sounds, with a leftward dominance for fast temporal modulations and/or a rightward dominance for slow temporal modulations. For example, previous studies from our laboratory have shown in humans that the left auditory cortex perceives brief acoustic events better than the right, but has a lower frequency resolution (Liégeois-Chauvel, Giraud, Badier, Marquis, & Chauvel, 2001; Liégeois-Chauvel, de Graaf, Laguitton, & Chauvel, 1999). Furthermore, using non linguistic stimuli that varied independently and parametrically in temporal or spectral dimension, two similar studies, with positron emission tomography (Zatorre & Belin, 2001) and functional magnetic resonance imaging (Jamison, Watkins, Bishop, & Matthews, 2006), revealed that activity in the primary auditory cortex followed the increase in temporal variation of the acoustic signal, whereas at a more associative level, in the anterior Superior Temporal Gyrus, activity followed the increase in spectral variation. From these bilateral activities, they further extracted an asymmetry index, lateralized to the left for temporal variations and to the right for spectral variations. More recently, capitalizing on the previously introduced spectrotemporal framework, Flinker and colleagues used a filtering technique first developed by Eliott and Theunissen to create speech stimuli with various degrees of temporal and spectral modulations (Elliott & Theunissen, 2009; Flinker et al., 2019). Using behavioral investigation, magnetoencephalographic and intracranial recordings they demonstrated differential involvement of both hemispheres during speech perception, with a left lateralization for temporal modulations and a weaker right lateralization for spectral modulations. In a similar approach, Albouy and colleagues created a corpus of ten sentences which were associated with ten different melodies resulting in an orthogonalized set of hundred sentences (Albouy et al., 2020). They then selectively filtered temporal or spectral modulations in the sung speech stimuli set and presented them in pairs while asking participants to report either whether the stimuli were the same sentences or whether the melodies were the same. Behavioral results indicated that degradation of temporal information resulted in poorer speech perception only whereas spectral filtering led to a decreased perception of melodies. Furthermore, they recorded brain responses using functional magnetic resonance imaging and showed that decoding of speech and melodies depends on patterns of activity in the left and right auditory regions respectively. This asymmetry was supported by specific sensitivity to spectrotemporal modulation rates within each region.

To account for experimental results at the time, Zatorre and colleagues propose a ***spectro-temporal trade-off*** model (Zatorre, Belin, & Penhune, 2002), to explain the asymmetry in hemispheric preference between language and music, based on different temporal and frequency resolution on the left and right auditory cortices. In a linear model the frequency and temporal aspects are intimately related according to the Acoustic Uncertainty Principle (Joos, 1948). In Fourier transforms, or in time-frequency analyses, the dilemma is always present between optimizing frequency and time resolution. Thus, and according to Zatorre and colleagues, to meet different computational constraints arising from environmental signals carrying information in the spectral and temporal domain, the brain would have developed two parallel and complementary networks, one in each hemisphere,

optimized in fast temporal or fine spectral processing, respectively. The functional asymmetry observed would thus correspond to a low-level computational specialization, not specific to any particular type of acoustic signal.

At the same time, David Poeppel proposed a related theoretical framework: the ***asymmetric asymmetric sampling theory*** (Poeppel, 2003, 2001). It is based on the fact that the speech signal contains (at least) two different but crucial timescales that must be equally processed for comprehension. One is associated with the syllables (150-300 ms, 4-7 Hz, theta rhythm) and the other associated with formantic transitions, characterizing the phonemes (20-40 ms, 25-40 Hz, gamma rhythm). He postulated a similar temporal organization within the auditory brain system. On the fact that the brain must discretize the continuous incoming sensory information. He postulated that auditory systems capitalise on (at least) two complementary mechanisms to chunk auditory information, each characterized by a specific temporal resolution. The central idea is that both local-to-global and global-to-local types of analyses are carried out concurrently (multitime-resolution processing). The brain is composed of different types of nerve cells that exhibit spontaneous oscillatory activity at varying rates. In particular, within the auditory system, there are large pyramidal cells with fast spontaneous oscillatory activity of about 40 Hz (integration constant of ~25 ms; gamma oscillations), and smaller cells with slow activity (integration constant of ~200 ms; 4-7 Hz, theta oscillations; (Hutsler & Galuske, 2003). The proportion of these cells in the right and left auditory cortex is furthermore asymmetric: the right auditory cortex is predominantly composed of small cells oscillating in the theta rhythm, whereas the left contains a greater proportion of large cells oscillating in the gamma frequency band. The theoretical model thus suggests the existence of a close coupling between the rhythmic activity of neurons in the left and right auditory cortex and the rhythm of the speech signal. More precisely, the left auditory cortex would rather provide a mode of division of information into small units, the phonemes, whereas the right auditory cortex would sample the signal in larger units and would thus be involved in the processing of syllabic information. Phonemic and syllabic information, processed in parallel by the left and right auditory cortex respectively, would then have the right format to be associated with linguistic representations stored in memory.

Despite the appeal of the theories in postulating a strong left-right dichotomy, it is more likely that these accounts are too reductionist and simplistic. Another theory suggests that the right hemisphere would be highly selective for particular acoustic cues, whereas the left hemisphere would be specialized for domain-specific mechanisms related to speech and not general purpose acoustic analyses (McGettigan & Scott, 2012; Scott & McGettigan, 2013).

In sum, there is substantial evidence for hemispheric asymmetries of the auditory cortical regions. These asymmetries are both of structural and functional nature. However, multiple questions remain concerning functional hemispheric asymmetry during speech processing. For instance, the extent to which it appears at the level of the first level of cortical processing (in Heschl's gyrus) or whether the asymmetry occurs further along the auditory

cortical hierarchy is still not clear. Moreover, what oscillatory regimes are characteristic of each hemisphere. Theoretical models postulate the existence of different neural populations with preferred time constant being distributed non uniformly across left and right hemispheres. This architectural constraint is thought to result in different oscillatory activity during speech processing. So far, the exact nature of these oscillations is still missing. Potential candidates have been proposed such as theta regime in the right hemisphere and gamma in the left. It is also highly probable that the spectral profile of both hemispheres' oscillatory activity contains multiple components with characteristics compatible with speech temporal dynamics. A critical question relates also to the type of processes at play during speech comprehension and whether as postulated by theoretical frameworks, a two-time-scales parallel processing is relevant or whether more processes at other time scales are needed to account for understanding.

Finally, it should be emphasized that the above-mentioned differences relate to trends, a relative specialization of the hemispheres. This phenomenon is not binary (zero or all) but there is a gradient. Concerning the language domain: stronger lateralization seems to be associated with better performance in verbal abilities (Barth, Boles, Giattina, & Penn, 2012; Boles, Barth, & Merrill, 2008; Catani et al., 2007; Chiarello, Welcome, Halderman, & Leonard, 2009; van Ettinger-Veenstra et al., 2010; Everts et al., 2009). Stronger functional cerebral asymmetry results in better (or poorer) performance independently of the hemisphere considered (higher left- or right-hemispheric lateralization). This is characterized by a u-shaped curve, having its midpoint close to a lateralization degree of zero and has been demonstrated repeatedly (Hirnstein, Hugdahl, & Hausmann, 2014). This implies that it is the degree of lateralization that is more important for performance than direction (e.g., whether a function is lateralized to the left or right hemisphere). The functional lateralization depends also heavily on specific contextual demands. This is an example of the flexibility with which the recruitment of specialized networks in our brain leads to successful complex cognitive processing which depends on both hemispheric specialization and interhemispheric interactions.

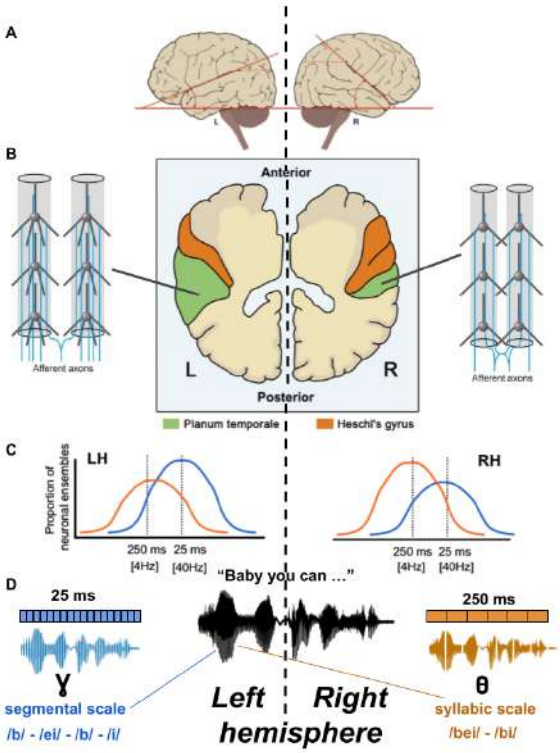


Figure 11. A) Structural asymmetries of the Sylvian Fissure. The left Sylvian fissure is longer than the right and runs at a more horizontal angle than the right. B) Structural asymmetries in the planum temporale and Heschl's gyri. The planum temporale has a larger volume in the left hemisphere than in the right hemisphere. In contrast, Heschl's gyri show a rightward asymmetry, with many individuals having two gyri in the right hemisphere, but only one in the left. Schematic depiction of two microcolumns in left and right hemispheric Planum Temporale. The density of dendrites and axons is higher on the left. In addition, left Planum Temporale neurons have a higher degree of arborization. Microcolumns are wider and further apart in the left hemisphere and afferents axons innervate smaller numbers of neighboring microcolumns, possibly enabling sharper tonotopic mapping of columnar frequencies. The higher density of dendrites and afferents on the left side could enable near-synchronous activation of frequency-specific microcolumnar neurons, thereby decreasing the latency of left Planum Temporale cells and increasing their temporal precision. C) Proposed mechanisms for asymmetric speech parsing: the left auditory cortex (LH) contains a larger proportion of neurons able to oscillate at gamma frequency than the right one (RH). D) Temporal relationship between the speech waveform and the two proposed integration timescales (in milliseconds) and associated brain rhythms (in hertz) according to the Asymmetric sampling in time hypothesis. Theta oscillations are thought to support the parsing of speech at the syllabic timescale while gamma oscillations enable chunking at the segmental scale. A, B Adapted from (Ocklenburg & Güntürkün, 2017); C adapted from (Giraud et al., 2007).

iii. On the possible origins of the asymmetry

Most biological systems show some degree of asymmetry (Geschwind & Galaburda, 1985). Indeed, accumulating experimental evidence attests the presence of brain asymmetry in multiple and distinct animal species (Bisazza, Rogers, & Vallortigara, 1998; Ocklenburg & Güntürkün, 2017). For instance, zebrafishes, rodents and also non-human primates show aspects of hemispheric asymmetry of brain function. These observations suggest that asymmetrical brain organisation is an evolutionary conserved design principle which according to some authors would grant various species with computational advantages

(Hartwigsen, Bengio, & Bzdok, 2021). For instance, asymmetry has been shown to be adaptive for chimpanzees, as individuals with higher hand preference for termite fishing were more efficient in this task (McGrew & Marchant, 1999). Increased visual asymmetry in pigeons and chicks is linked to higher ability in discriminating grains from pebbles (Güntürkün et al., 2000; Rogers, Zucca, & Vallortigara, 2004) (Figure 12).

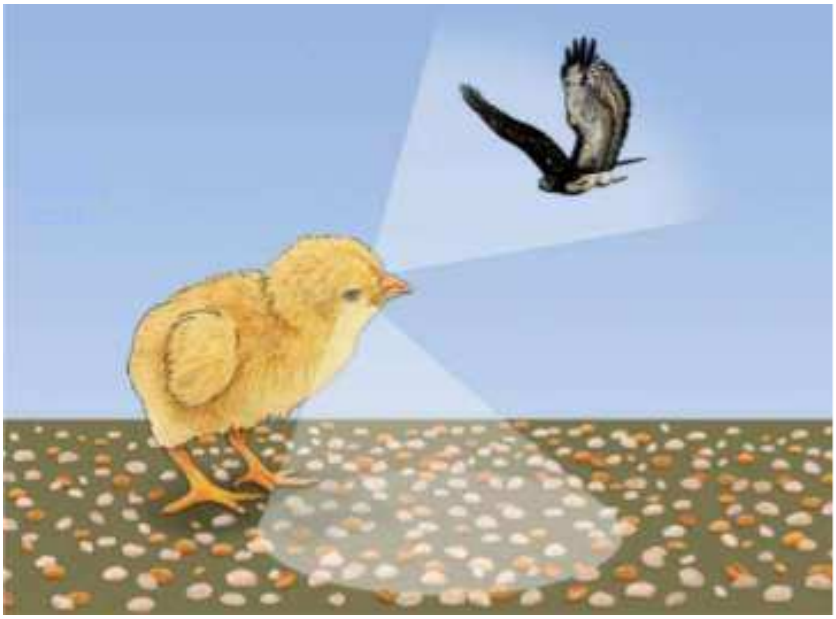


Figure 12. Brain lateralization in domestic chicks is associated with an advanced ability to perform two tasks in parallel. In this study chicks had to find grains among pebbles and were simultaneously forced to be vigilant for birds of prey. Lateralized birds could do this well, while non lateralized ones failed. Adapted from (Güntürkün, Ströckens, & Ocklenburg, 2020).

Several explanations for the emergence of hemispheric asymmetries have been postulated and are related to theoretical advantages. First of all, while the brain only represents about 2% of the body mass of an adult human, it accounts for about 20% of the total body calories consumption (Raichle & Gusnard, 2002). As a result, a strong evolutionary pressure must have been at play to use brain tissue as efficiently as possible. Hemispheric specialization might have been a good solution to increase neural capacity in compartmenting a specific cognitive function to a hemisphere thus avoiding its unnecessary duplication in the other hemisphere (Vallortigara, 2006). Another evolutionary pressure is the fact the human skull is limited in size by the restrictions of the birth canal, as such hemispheric specialization, by making an efficient use of neural tissue might have been a particularly good answer to that constraint. Apart from the maximization of available space by reducing redundancy between brain regions, hemispheric specialization may also contribute to an increase in parallel processing by enabling separate computations of complementary information into the two hemispheres. Being able to perform different computations at the same time has enabled an increasing degree of cognitive sophistication that might be the basis for our highly complex cognitive ability to process and comprehend language with apparently little effort. Moreover, hemispherically divergent processing may prevent response competition between both

hemispheres. Furthermore, it is believed that hemispheric specialization allows higher processing speed and efficiency as computation may be performed within only one hemisphere, reducing the reliance on interhemispheric information transfer via the corpus callosum (Ringo, Doty, Demeter, & Simard, 1994). This advantage may stem from evolutionary pressure in imposing time constraints in information transfer. In sum, asymmetry would be advantageous for biological organisms as part of their neurocomputational infrastructure as it would allow increasing task performance while decreasing reaction times, and also it would allow the possibility to run parallel and complementary neural processes.

Today, it is assumed that most species differences in lateralization (language asymmetrical organization included) are determined by complex and interacting genetic and non-genetic influences such as local ecological adaptations (Esteves, Lopes, Almeida, Sousa, & Leite-Almeida, 2020; Ocklenburg & Güntürkün, 2017). As an illustration that genetic factors do not constitute a full explanation of functional lateralization, a twin study indicated that only 25% of the variance in handedness data was accounted for by genetic effects, while the 75% remaining were determined by non-shared environmental influence (Medland et al., 2009). With regard to language processing, there is convincing evidence for a lateralization of production and perception of conspecific vocalizations in the left hemisphere in various species of primates (Ocklenburg & Güntürkün, 2012). However, it seems that humans have dramatically expanded this asymmetry during the course of language evolution. Such observations have led researchers to believe that we inherited our left hemispheric dominance for language processing from our ancestors and then further develop our unique language system on this common mammalian pattern. This thesis is postulated by Ocklenburg and Güntürkün who proposed a phylogenetically early emergence of speech communication asymmetries, according to which human language lateralization evolved due to a left-hemisphere dominance for more fundamental features of species-typical communicative sounds or their production, rather than due to a left-hemisphere dominance of language itself (Ocklenburg & Güntürkün, 2012). Other theories have assumed a common genetic determinant for language and motor lateralization as data demonstrated a high prevalence of concordant hand preference and hemispheric dominance in language perception (Bethmann, Tempelmann, De Bleser, Scheich, & Brechmann, 2007). In this view, speech is envisaged as a special form of motor cognition relying on the dorsal stream and auditory-mapping mechanisms which is coherent with the left-lateralized dominance in motor cognition (Annett, 1998).

Although functional asymmetries exist, both hemispheres closely interact during many cognitive operations and brain lateralization may be more appropriately viewed of degree, not of kind. As such, it is very typical that both hemispheres have some ability to perform the same task, with one hemisphere generally being more efficient than the other. This suggests that the two hemispheres may play somewhat complementary roles in the intact brain, given the nature of their contributions.

2.4 Toward integrated models of speech understanding

i. auditory implication in speech processing

The adequacy between the rhythms present in the speech signal and the periodic neural activity present at rest in the auditory cortices are at the basis of the emergence of oscillatory theories of language processing (Giraud et al., 2007; Keitel & Gross, 2016; Lubinus et al., 2021; Morillon et al., 2012). These theoretical models propose that low gamma (25-35 Hz), theta (4-8 Hz), and delta (1-3 Hz) oscillations play a role at the phonemic, syllabic, and phrase level processing, respectively, during speech perception (Giraud & Poeppel, 2012). In particular, the authors suggested the existence of a cascade of processes that would allow the transformation of the continuous speech signal into distinct units and that would be based on a close coupling or nesting between theta and gamma oscillations. After encoding the spectro-temporal properties of the speech signal in the primary auditory cortex, salient information (edges or peaks within the amplitude envelope) would induce a phase reset of the intrinsic theta and probably gamma oscillations (step 1). Theta oscillations would specifically align with the slow modulations of the amplitude envelope that characterize speech rate in which syllabic information dominates (step 2). Theta and gamma oscillations, weakly coupled at rest and processing verbal information in parallel, would then become strongly coupled: the phase of theta oscillations would modulate the amplitude (and potentially the phase) of gamma oscillations (theta-gamma nesting; step 3). The power of the gamma oscillations would then regulate the neuronal discharge from the primary auditory cortex to higher level areas to create excitability cycles corresponding to the phonemic and syllabic scales (step 4). Finally, neural excitability would be modulated to be aligned with the structure of the speech signal (step 5). The authors thus propose that the construction of intelligible speech representations relies on both (1) phase-locking between incoming information and auditory cortex on at least two time scales, and (2) hierarchical coupling between theta and gamma cortical oscillations that would allow phonemic and syllabic processing to be performed in parallel while remaining hierarchical. Thus, the speech signal would reset the endogenous oscillatory activity of the auditory cortex, mainly in the frequency bands corresponding to the optimal sampling rates for phonemic and syllabic parsing. In return, the cortical oscillations would allow the temporal organization of the continuous speech signal to facilitate its recognition (Giraud & Poeppel, 2012).

Specific hypotheses postulated by the proposed oscillatory model have received experimental support. Indeed during speech processing, the coupling between the amplitude of the stimulus envelope and the phase of the low frequency neural activity has been repeatedly observed (Kösem et al., 2018; Luo & Poeppel, 2007; Peelle & Davis, 2012). Coupling between theta and gamma oscillations in auditory cortex has also been reported (Gross et al., 2013; Morillon et al., 2012). However, to date compelling evidence in favor of a sampling of phonemic information at the gamma rhythm is still missing. Moreover, the model does not

explicitly describe how internal knowledge of language (speech sounds, linguistic structure etc...) guides perception.

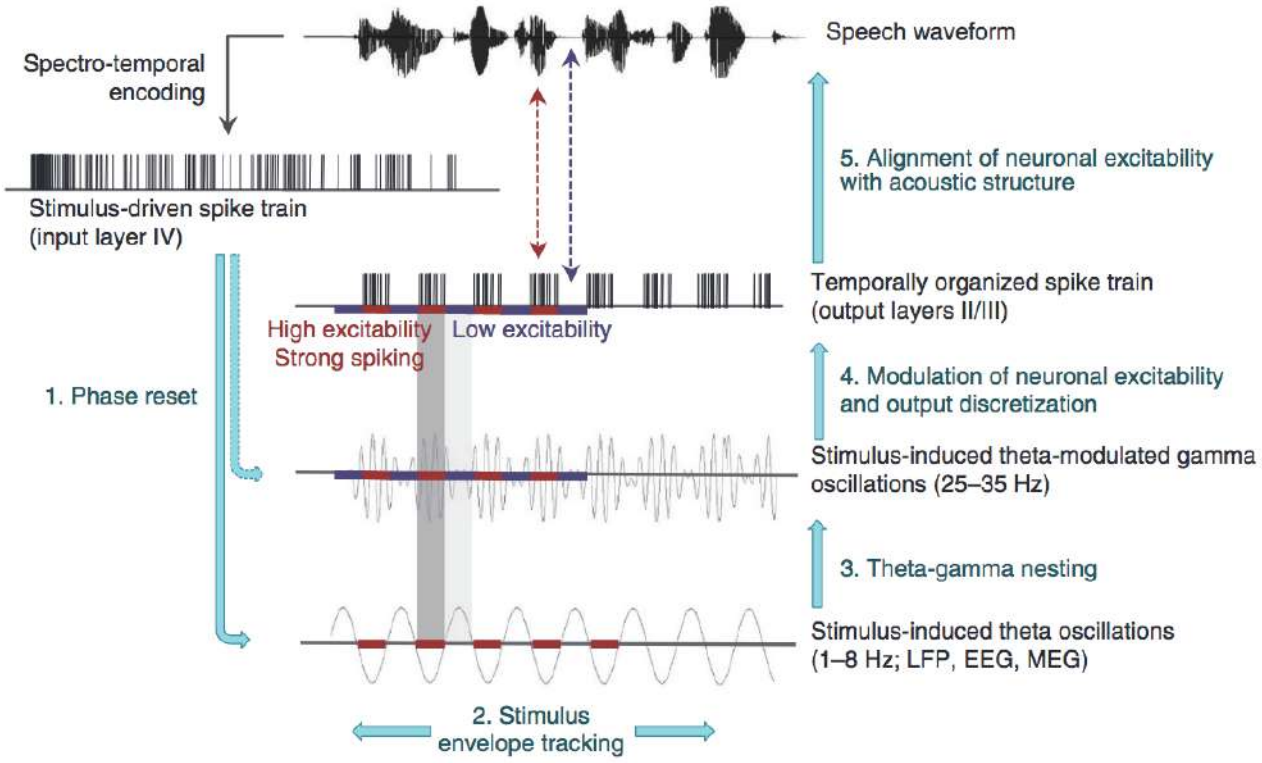


Figure 13. Presentation of an oscillation-based model of early computations at play during speech perception. Schematic representation of the five operations allowing connected speech to be parsed by cortical theta and gamma oscillations. Adapted from (Giraud & Poeppel, 2012).

ii. Motor implication in speech processing

There is evidence for the presence of beta oscillatory activity in sensorimotor areas during temporal tracking as well as during continuous speech processing (Arnal, Doelling, & Poeppel, 2015; Morillon & Baillet, 2017; Saleh, Reimer, Penn, Ojakangas, & Hatsopoulos, 2010). Experimental observations support the fact that beta-band activity in motor regions modulates auditory processes through predictions of incoming sensory information. This top-down mechanism which enable to efficiently allocate processing resources, by predicting upcoming inputs and reducing uncertainty leading to an enhanced processing of noisy inputs is at the basis of the ‘predictive routing’ account (Spitzer & Haegens, 2017). It is thought to result in an anticipatory phase reset, coming from higher-level processes and not directly driven by low-level stimulus features, of ongoing oscillations in brain areas involved in the

processing of the expected sensory event (beat or syllables) (Barczak et al., 2018). The contribution of motor areas to the analysis of perceptual temporal dynamics is probably modality- and domain-general and is thought to occur during the perception of any sensory signal that is temporally organized at low frequency. This mechanism has been demonstrated during speech processing. For instance, Park and colleagues showed that activity predominantly in the left frontal motor regions modulated the phase of delta and theta oscillations in the left auditory cortex when listening to intelligible speech compared to unintelligible (backward) speech (Park, Ince, Schyns, Thut, & Gross, 2015). The stronger these top-down influences were, the better the brain-speech alignment. Motor delta oscillations have been shown to impose temporal constraints on sensory information sampling. As such, they control perceptual sampling by providing a temporal frame that both amplifies and orders sensory processing. They also optimize the parsing, encoding and processing of slow linguistic information (reviewed in Morillon, Arnal, Schroeder, & Keitel, 2019). Moreover, optimal audiomotor coupling is observed during the presentation of stimuli within the range of the natural syllabic rate (~4 Hz) and during speech perception (Assaneo & Poeppel, 2018; Keitel, Ince, Gross, & Kayser, 2017; Morillon et al., 2010; Park et al., 2015). In cortical motor areas, both delta-tracking of the phrasal acoustic rate and delta-beta coupling predicts speech comprehension (Keitel, Gross, & Kayser, 2018).

The influence of frontal motor regions on auditory cortex is consistent with dual-pathway models (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009) and the idea that speech production regions are able to make predictions about the perceived signal in a way that facilitates phonemic categorisation and thus comprehension (Iacoboni, 2008). The motor cortex could thus control (monitor) the rhythm of auditory verbal information (related to the rhythmic articulatory movements necessary for production) in order to better anticipate the occurrence of acoustic events relevant for recognition (Arnal et al., 2015; Morillon & Baillet, 2017; Morillon, Schroeder, & Wyart, 2014; Morillon & Schroeder, 2015; Strauß & Schwartz, 2017). Speech motor areas may in addition represent phonological information (Du & Zatorre, 2017).

It is now becoming increasingly clear that speech comprehension relies on a tight relationship between the perceptual and the motor system (Schwartz et al., 2012; Skipper, van Wassenhove, Nusbaum, & Small, 2007). However, computational models that integrate motor and perceptual processes into one cognitive speech communication system are far from being detailed at the neural implementational level. The role played by the motor cortex in speech perception is still actively debated (Skipper, Devlin, & Lametti, 2017). This is due in part to the under characterization of the mechanisms and pathways linking auditory and motor areas. Indeed, while data demonstrate a strong interconnection between the primary motor, somatosensory, and auditory cortex seem (Skipper & Hasson, 2017), there is still a lack of evidence demonstrating direct projections between motor and auditory regions in humans (Morillon, Hackett, Kajikawa, & Schroeder, 2015). Finally, the origins of top-down linguistic predictions, being at the phonological, lexical or sentence level are highly complex and poorly understood, but audio motor coupling may be a key mechanism to properly understand the motor contributions in speech comprehension (Strauß & Schwartz, 2017).

iii. Integrated view of speech processing

Building on the 'analysis-by-synthesis' framework (Kuhl et al., 2014; Liberman et al., 1967), Martin proposed to envisaged language comprehension as perceptual inference in which cues from different nature and origin are weighted and integrated to derive meaning (A. E. Martin, 2016, 2020). In this view, signal-based cues which are related to acoustical qualities of the speech signal (including voice onset time, formant values, speech rate, etc...) are combined and integrated with knowledge-based cues representing statistical and structural knowledge of language learned through experience and stored in memory, such as semantic or lexical knowledge to robustly decode speech. The inferred linguistic units (phonemes, syllables, words, sentences, and higher-level structures) can, in turn, act as endogenous cues for further processing. The resulting iterative processes enable the building of language comprehension across all levels of linguistic hierarchy. The model predicts a hierarchy of temporal receptive windows from early sensory to higher-level brain areas, consistent with neurophysiological evidence (Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Lerner, Honey, Silbert, & Hasson, 2011). This framework adds another level of complexity to previous oscillatory based models by completing them with an internal linguistic model. A computational model which implements such principles has been shown to be optimally sensitive to the natural temporal speech dynamics and even explain empirical data on speech illusions (Ten Oever & Martin, 2021). Finally, Kaufeld and colleagues have recently shown that oscillatory activity during speech perception reflects not only processing of acoustic properties of the speech signal but also its linguistic structure and meaning (Kaufeld et al., 2020).

The set of computations at play during such a sophisticated task like speech understanding is likely to be numerous, complex, hierarchical and at some degree, happening in parallel. For these reasons, it is difficult to sketch a complete picture of both the multiple underlying computation mechanisms and the neural architecture supporting them. However, tremendous efforts have been made in the last decades to characterize into more details this truly human feat and to get integrated models of (spoken) language comprehension. Based on the integration of experimental data from various sources, key mechanisms and principles are starting to emerge. Oscillatory models of speech processing have been instrumental in the description of the multiple computations needed to transform an acoustic waveform into discrete linguistic representations. However, while bottom-up processes are required for speech comprehension, top-down processes are at least equally important in this regard. Experimental works support the fact that such mechanisms could originate from higher order areas including cortical motor regions. Indeed, the high interplay between motor and auditory regions is now recognized as a sine qua non condition on which language comprehension is built. While details are still missing, current integrated models largely include motor components. The precise nature of their function is still debated, and more work is needed to clarify it and integrate it in a general framework. Current view on language processing has moved away from a classic and static model, therefore a better characterisation of the links

and interactions between the multiple cortical regions but also mechanisms are necessary to finally get a chance to deliver a computational(ist) neurobiology of language. In this quest, neuroscientists, linguists and psycho-linguists alike can rely on the rapid technological development while keeping in mind that careful behavioral investigation of human perceptual ability can play an important part to solve the language question.

3. Methodological considerations for studying speech processing in Human

Neuroscientists studying language processing must face multiple challenges. For instance, animal models can only afford relative answers to the bigger question. Moreover, studying human subjects implies being limited by methodological constraints such as non-invasive measurements. However, such challenges can turn into opportunities to develop clever behavioral paradigms and new ways of analyzing and looking at the data. Indeed, an ever-increasing number of new technologies have made the manipulation of auditory signals much faster and easier. In combination, the development of new non-invasive recording techniques as well as sophisticated data analysis methods have paved the way to exciting discoveries in human auditory and speech processing (Baillet, 2017; Huth et al., 2016; Richards et al., 2019). Although there has been quite some optimism surrounding such a technical development, some authors have advocated for a careful examination of behavior as a means of providing understanding of complex cognitive function (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). They have argued that study of the neural implementation of behavior can highly benefit from prior behavioral work.

3.1 Behavioral approaches

i. Experimental paradigms

The experimental methodology rests on the careful investigation of the relationship between behavioral patterns and stimuli properties in a controlled environment. Although it can be seen as being in contrast to speech perception in everyday situations, multiple experimental paradigms with diverse degrees of naturalness exist and form the basis for speech comprehension investigation. The study of speech and language processing has considerably evolved over time thanks to methodological advances but also scientific practices. Indeed nowadays, more and more data and software are instantaneously available and free to the scientific community for reuse and analysis. Creative and innovative behavioral paradigms have been developed and they have been instrumental in the development of auditory perception scientific knowledge. Although there exists a myriad of ways of investigating auditory and speech perception, I will focus only on two relevant behavioral paradigms for the scientific questions investigated in this PhD work. They concern ways to highlight hemispheric differences in processing speech material but also ways to probe online mechanisms at play during spoken language comprehension.

Sixty years ago, Doreen Kimura introduced ***the dichotic listening paradigm*** (Kimura, 1961). Since then, it has been widely used in auditory and speech research in both healthy and pathological populations. It allows for the assessment of hemispheric lateralization in a simple and non-invasive manner. The test involves the simultaneous presentation of two different stimuli, one in each ear. The stimuli used are usually pairs of syllables, words, numbers or even sentences. Participants are usually asked to report the stimulus that they hear best; normally this repetition results in an ear advantage, as stimuli are usually better recognised in one ear than in the other. In healthy subjects, Kimura showed that the majority of right-handed people are better at identifying stimuli presented in the right ear than in the left ear when these are verbal in nature: this is known as the right-ear advantage. This right-ear advantage reflects the representation of language in the left hemisphere. The identification of nonword sounds such as humming, coughing or laughing did not result in a right-ear superiority, suggesting that verbal nature was critical for this effect (King & Kimura, 1972). This right ear advantage has been claimed to result either from structural constraints or to attentional mechanisms. In any case, the behavioral performance in the dichotic test and result from the Wada test have been shown to be highly concordant (Hugdahl, Carlsson, Uvebrant, & Lundervold, 1997; Zatorre, 1989). Finally, the right ear advantage in verbal dichotic listening is a global phenomenon as it has been reported across various languages and language families (Bless et al., 2015).

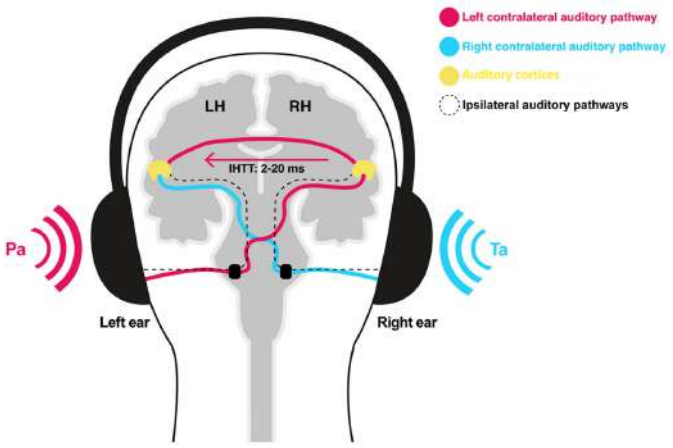


Figure 14. During dichotic listening, one syllable is presented to the left ear (PA) and the other one is simultaneously presented to the right ear (TA). Both ears are connected anatomically via ipsilateral (black dotted lines) and contralateral pathways with (pink and cyan lines) the auditory cortices. The cyan line indicates the contralateral pathway transmitting the right ear stimulus directly to the speech dominant left hemisphere. In contrast, the syllable running from the left ear to the right hemisphere has to cross the corpus callosum to be efficiently processed (pink line). Yellow: Right and left primary and secondary auditory cortices. LH: Left hemisphere, RH: Right hemisphere. IHTT: Interhemispheric transfer time. Adapted from (Steinmann, Leicht, Andreou, Polomac, & Mulert, 2017).

The gating paradigm was first introduced by Grosjean in 1980 (Grosjean, 1980). It has been used many times in spoken word recognition research. In this behavioral task, participants are presented with segments of spoken material (words, sentences) of increasing

duration. After each stimulus auditory segment presentation, participants are asked to guess the stimulus presented and to give confidence rating on their response. These successive presentations of longer chunks of auditory stimuli are called gates. Their number can vary across studies depending on the research question but are fixed within an experiment. The first segment is usually very short and the last one corresponds to the entire stimulus. At each trial corresponding to a particular stimulus but also to a specific number of gates, an isolation point can be retrieved; it corresponds to the gate at which the participant gave the correct response (she identified correctly the stimulus presented) and did not change her response hereafter. It determines the size of the segment (in ms) needed to recognize the stimulus. This specific presentation scheme allows a tight control over the acoustic-phonetic information presented to subjects at each gate. This paradigm has been fruitful in revealing and confirming robust effects of various linguistic variables including word frequency, length and amount of context on speech recognition. Previous studies demonstrated that less acoustic-phonetic information is needed for short, frequent and in context words than for long, infrequent and isolated words (Grosjean, 1980). The second experimental contribution of this PhD work includes a variant of the gating paradigm to assess the contribution of multiple variables on speech comprehension. In our own and newly developed version of the paradigm, by presenting successively decreasing time-compressed versions of spoken utterances and probing participants guesses at each gate, we demonstrated the individual influence of multilevel linguistic features on behavioral performance.

ii. Nature of the auditory stimuli

Past experimental studies on auditory and speech processing have relied on different kinds of auditory stimuli. Traditionally, seminal studies used highly controlled experimental material to gain insight in basic and critical auditory processing mechanisms. Moreover, such investigations have often focused on one specific characteristic (physical, linguistic) at a time. The construction of stimuli is thus crucial to successfully answer a research question. Complexity and naturalness are two really important criteria to take into account when planning experiments. Pure tones for instance have been largely used by the auditory research community, such synthetic stimuli can be built in a parametrically manner, controlling every acoustical aspect. It is then possible to induce systematic variations along a particular dimension while holding all other parameters constant and assess the behavioral or neural relevance of such manipulation. At present, the availability of a large set of open access and free software makes the creation of tightly controlled materials easy and fast. However, a concern with such stimuli is their ecological validity. Researchers must be aware of the limitations of such experimental approaches. A step further on the naturalness spectrum, is the use of real sounds (environmental, music or speech). This approach has been endorsed by a large number of researchers in the field. Usually, isolated words matched for on diverse dimensions (number of syllables, frequency of occurrence, etc...) are part of the experimental

conditions in which participants can be asked to perform discrimination or lexical decision task. This approach has been highly successful and fruitful in describing robust effects that are part of the “textbook view” in speech processing. As the field and the questions evolved and expanded, longer and more complex materials were developed and used. Studies based on isolated sentences are now widely represented in the literature. They allow the investigation of processing steps of higher complexity such as syntax processing for instance. While it is arguably more ecologically valid than isolated words, in everyday life we are rarely confronted with sentences out of context. Overcoming such challenges, recently the idea of studying speech and language processing in a more ecological way has been pushed forward (Hamilton & Huth, 2020). In this view, multiple studies have employed long excerpts of real speech drawn from real-world sources including audiobooks, radio shows or podcasts. By virtue of their nature, these narratives include rich contextual aspects which are typically encountered in real life. This complexity and naturalness enable the fine-grained characterization of multiple different linguistic levels and multiple timescales, in combination with diverse linguistic and extra-linguistic processes at play during comprehension. This has the potential to bring new insights into theoretically postulated neural mechanisms such as predictive coding and the concept of temporal receptive windows of varying lengths for instance. Another advantage is that studies that include this specific type of stimuli produce results that are more easily generalisable to everyday language use. Additionally, an important point is that the datasets that are generated during the course of these studies (neurophysiological recordings for instance) can be reused and analyzed with a different angle and hypothesis by different teams. This is in stark contrast to previous approaches and can greatly increase access to meaningful data while reducing the time between hypothesis development and hypothesis testing. Although the use of more naturalistic experimental material might be highly beneficial to the research field, it must be noted that it also brings new challenges that must be overcome. Indeed, experimenters have far less control over audio excerpts from audiobooks for instance. High correlations over various variables of interest may limit interpretations of the results. Because of the complexity of the stimuli, researchers may adopt sophisticated data analysis, or modelling approaches resulting in less clear-cut conclusions than standard analysis strategies.

In sum, when studying speech and language perception, researchers have a large panoply of auditory stimuli they can use. The final choice must be based on the research question at hand. Since each category of experimental material has advantages but also disadvantages, it is really important to consider them in the context of the hypothesis being tested. Occasionally, combining different types of stimuli can be highly relevant to answer specific questions. Sometimes a compromise between naturalness and artificialness has to be found in order to have sufficient experimental material while avoiding long and painful recording sessions with professional actors for instance. New technologies like text-to-speech algorithms can be highly relevant in this situation.

iii. Artificial modifications of sounds

Clever manipulation of sounds and auditory stimuli have enabled tremendous progress in our understanding of the mechanisms that support speech and language processing. This has been rendered possible thanks to the development of simple but also highly sophisticated tools dedicated to sound manipulations. Many are now an integral part of the toolboxes available to researchers to ask questions concerning speech or sound perception. Simple modification techniques include adding noise to reduce signal-to-noise ratio for example. A large number of studies have used noise with various characteristics (from broadband noise to sounds with more complex characteristics to speech) to degrade speech material. In his seminal studies in the 1950s, Cherry integrated speech into a simultaneous speech stream to assess the limitations of the speech processing system in what he called the ***cocktail party*** phenomenon (Cherry, 1953). Later on, and thanks to the evolution of tools to modify sounds, researchers have started to use various filtering techniques to remove either specific temporal or spectral features of the speech sounds and evaluate the behavioral consequences of these modifications. One of the most popular methods is the ***vocoder*** which has already been introduced in previous sections. It was invented in 1938 by Homer Dudley at Bell Labs. It consists in the decomposition of the speech waveform into different filtered versions of the signal in frequency bands thanks to the use of a model of the cochlea processing. Following such decomposition, it is possible to remove any filtered version before resynthesizing the speech waveform. Such modifications result in decreasing the frequency details of the signal while mostly keeping the temporal information. The main findings using this technology is that humans can adapt to heavily frequency degraded speech (Shannon et al., 1995). On the other hand, studies have filtered the speech signal in the temporal domain by either applying low-pass filters or even band-pass filters to target specific temporal fluctuations. They have been instrumental in implicating the low-frequency temporal modulations for speech comprehension (Doelling et al., 2014; Ghitza, 2012). Filtering methods have been applied in conjunction with the dichotic listening paradigm to assess the implication of each hemisphere in processing different time scales present in the speech signal (Chait, Greenberg, Arai, Simon, & Poeppel, 2015; Saoud et al., 2012). More recently, experimental works using the spectrotemporal framework introduced in a previous section (1.1.iii), have been able to describe in a detailed way the spectrotemporal scales needed for speech understanding. Of note, different spectrotemporal features are important to distinguish between different types of acoustic signals (speech and music for example) (Albouy et al., 2020; Flinker et al., 2019). ***Sine Wave speech*** is a radical transformation applied to speech signals, in which the formants of syllables are replaced with frequency-modulated sine waves that follow their center frequency. At first the sounds created do not at all resemble any audible speech as attested by the fact that listeners process this modified signal in distinctive ways depending on whether or not they are told that it is speech. Once listeners have been instructed to process the stimuli as speech, they cannot switch back to a non-speech mode of processing the stimuli (Best, Morrongiello, & Robson, 1981; Remez et al., 1981). These studies demonstrate the reliance on top-down information during speech comprehension. By

selectively manipulating speech fine structures and enveloppe, Smith and colleagues created ***auditory chimaeras*** (Smith, Delgutte, & Oxenham, 2002). These modified acoustic signals contain the envelope of one sound and the fine structure of another. It is then possible to assess the dependence of our perceptual system on these individual cues. In their study, they show that the envelope is most important for speech perception, while the fine structure is decisive for pitch perception and sound localization. Moreover, when the two features are conflicting, the localisation of the speech sound is determined by the fine structure, whereas the words are identified according to the envelope. Time stretching methods have also been employed to characterize the constraints of the perceptual system dedicated to speech. Various experiments used ***time compressed speech*** material and showed that our cognitive system can adapt to a certain extent. Indeed, as long as speech rate remains in a particular range of values, understanding is still possible, however when speech is presented at 2.5 times its natural speed, comprehension drops abruptly (Peelle et al., 2004). Pitch modification technique can also be used to examine the tolerance of the speech perception system to variations. Modifications of the temporal structure of the speech signal have also been used to investigate mechanisms underlying speech perception. For instance, locally time-reversed speech has been used to investigate intelligibility. The surprising result is that if a speech waveform is cut into very short successive segments (20 to 40 ms) and each segment is reversed in time, listeners do not have any problem understanding. However, when cut into segments of longer duration and locally reversed, intelligibility declines (Saberi & Perrott, 1999; Ueda, Nakajima, Ellermeier, & Kattner, 2017). At larger time scales, Lerner and colleagues have presented stories scrambled at various levels including words, sentences, and paragraphs (Lerner et al., 2011). They found that different brain areas were influenced by coherent information at different temporal windows (from very short timescales in the primary auditory cortex to longer ones in higher level regions). Using a related approach, Overath and colleagues constructed ***sound quilts*** by shuffling segments of natural sounds (Overath, McDermott, Zarate, & Poeppel, 2015). The modification resulted in sounds with mostly preserved properties on short timescales while they were disrupted on longer scales. Similar to Lerner and colleagues, they found that while primary auditory regions were not sensitive to such modifications, bilateral regions at higher levels in the auditory hierarchy (Superior Temporal Sulcus) show systematic modulations of their activity.

In conclusion, speech and sound artificial modifications are crucial to address multiple aspects of speech perception. They can foster the discovery of the mechanisms implicated in temporal, spectral or spectro temporal processing during speech perception.

3.2 Modelling approaches

The comprehension of mechanisms underlying speech processing can proceed in various ways. Behavioral and neurophysiological studies have been extremely useful and fruitful in this regard. There is now momentum to more explicitly describe the specific computations involved during language processing, and it is argued that the computational modeling approach can greatly assist in this endeavor while also helping refine current theoretical frameworks (Guest & Martin, 2021). Although there is a resurgence of these ideas, modeling approaches have long been used in the language and brain sciences. However, thanks to the exponential rise in computational resources and methodological tools, it is now more than ever, easy to use on a daily basis such tools to gain insight on comprehension mechanisms. In the area of speech and language comprehension, a variety of models have been proposed to account for experimental data. They are based on different architectures and seek to model and explain different aspects of the ability to use and understand language.

Oscillatory models For instance Hyafil and colleagues proposed a computational model with a neural inspired architecture with oscillatory mechanisms at its core (Hyafil, Fontolan, Kabdebon, Gutkin, & Giraud, 2015). Based on previous neurophysiological works, they implemented well known neural mechanisms including theta and gamma oscillatory activity into a neuro-inspired architecture thanks to networks of coupled oscillators made of excitatory and inhibitory populations of neurons. These networks exhibit complex oscillatory behaviors depending on specific parameters such as the relative proportion of inhibitory neurons versus excitatory neurons but also incoming inputs. They showed that such architecture could flexibly track the syllabic rhythm of speech and temporally organize the response of gamma neurons to decode both phonemes and syllables thanks to a specific coupling regime. As such, their model, thanks to oscillation-based computations, enables speech de-multiplexing, parsing, and encoding. Moreover, the advantage of their modelling work is the fact that it can achieve online syllable boundary detection as accurately as state-of-the-art offline syllable detection algorithms. In a refined version, Hovsepyan and colleagues, combined oscillatory based mechanisms and predictive coding and provided a neurocomputational architecture that was able to deal with the continuous nature of speech and the difficulty associated with its parsing into discrete representations, as reflected by its performance in online syllable detection (Hovsepyan, Olasagasti, & Giraud, 2020).

While the previous examples of modelling work focused mainly on the segmentation and decoding of the speech signal, a large proportion of work aimed at modelling higher level mechanisms implicated in spoken language recognition and comprehension.

The cohort model, for instance, attempts to account for how auditory input is mapped onto a word into the listener's lexicon (Marslen-Wilson & Welsh, 1978). It was the first psycholinguistic model of spoken word recognition. The online presentation of a specific word results in the real-time activation of every word stored in the lexicon which starts with the segments composing the input. By the time all segments from the word input are available to the model, every other word is ruled out, leaving the only word present in the lexicon that

matches the input. In this model, spoken-word recognition takes place thanks to three stages: access, selection, and integration. Acoustic–phonetic elements in the speech signal are mapped onto words in the lexicon during the access phase. The cohort then is constituted of all words that match with the input. In the selection phase, possible words mismatching the incoming speech by more than a single feature are removed. Finally, during integration, activated words' syntactic and semantic properties are retrieved and checked according to contextual constraints. However, the model has been challenged by a number of behavioral findings (Weber & Scharenborg, 2012). One of the limitations of the model is that it does not take into account context effects as it is based exclusively on the acoustic and phonetic characteristics of the input signal.

The Trace model is an interactive-activation, localist connectionist model of spoken word recognition (McClelland & Elman, 1986). It assumes three levels of primitive processing units or nodes. In the lowest level, the nodes represent phonetic features, then in the next, they represent phonetic segments, and finally, the last nodes represent words. Trace's processing units have excitatory and inhibitory connections between and among levels respectively. Such architecture enables the control of the activation level of the units depending both on the inputs and also the activity of the overall system. By allowing highly activated nodes to inhibit competitive nodes, it results in a "winner takes all" decision. Indeed, the nodes can be regarded as detectors of specific features, phonemes, or words. Activation is passed between levels, and used by the model to confirm and accentuate evidence for the presence of a given feature, phoneme, and word within the input. Moreover, all connections between levels are bi-directional, as such information can flow in both directions. This results in the occurrence of both bottom-up and top-down processing. The model has been successfully used for quite some time and has accounted for many aspects of speech perception, such as the perceptual phoneme restoration or the categorical perception for instance.

The Shortlist model, a descendant of Trace, is also a localist connectionist model of spoken word recognition (Norris, 1994). Shortlist is a feed-forward only model. It is composed of two stages. In the first stage, a "shortlist" of word candidates is activated based on a match between lexical item and the speech input. In the second stage of processing, the shortlist of lexical items enters into a network of word units which compete with one another via lateral inhibitory links for recognition. Shortlist B has been introduced as a newer version of the original Shortlist model. It is based on Bayesian principles (Norris & McQueen, 2008). It no longer uses activations of word candidates, but word probabilities. It also incorporates word frequencies as prior probabilities, and computes likelihoods as a means of handling mismatches in the input. The model successfully accounted for multiple behavioral findings, such as word frequency effects for instance (Dahan, Magnuson, & Tanenhaus, 2001).

More recently new types of models have emerged, they are based on the development of natural language processing methods. The underlying principle is that they model probabilistic properties of language thanks to large text corpora.

The n-gram model is one of the first models of this type; it is a Markov model which computes the probability of any word within a large corpus while taking into account the

preceding context (the previous words). The probability of a word is computed as the relative frequency of co-occurrence of word sequences of different length present in language corpora. For instance, a 3-gram model takes into account two preceding words to compute the conditional probability of occurrence of the third word. They can be used to model words but also syntactic categories of words (parts-of-speech). These models require huge amounts of linguistic data (large corpus) and estimating the conditional probability is time consuming and computationally intensive as the number of 'n' increases.

Vector based models The main idea is that words that tend to co-occur often together may have specific links between them. In these models, words are represented as numerical vectors of high dimensions (see Erk, 2012 for a review). As such every word has a specific position in high dimensional space. Words that are close in this space have related meaning. For instance, "king" would be closer to "queen" than "carrot". Moreover, as each word has its own coordinates, it is also possible to combine them through vector computations in order to explore the model language representation. For example, looking at the word with coordinates associated with the combination "king" - "man" + "woman", would result in the word vector "queen" (Mikolov, Yih, & Zweig, 2013). Neuroimaging studies have used such language representation to ask whether the brain could use a similar mechanism during language processing. Although these models capture some structural aspects of language, they lack precise contextual information (Huth et al., 2016).

Transformers Language Models are the latest generation of language models; they are composed of deep neural networks with attentional mechanisms. These models try to extract structure from languages through unsupervised learning. Usually, such models are trained on tasks such as word masking in which they are presented with written sentences which lack a word at a specific position in the sentence. The deep neural network goal is then to retrieve the missing word. They are able to perform the task thanks to large text corpora and extensive training phase. Their performance on various tasks are very high and they have been progressively used in experimental works. They are successful at predicting brain activity during natural speech processing (Caucheteux, Gramfort, & King, 2021; Goldstein et al., 2020; Heilbron, Armeni, Schoffelen, Hagoort, & de Lange, 2020; Schrimpf et al., 2020).

Modeling approaches are very promising in their ability to uncover the mechanisms of language processing. By formalizing theoretical points, they allow modelers to test hypotheses and refine theory. While they have obvious advantages, they bring new challenges and have limitations that will need to be overcome in order to expand our understanding of language processing. There are a variety of language models available with different architecture but also different goals. They vary along a continuum in terms of architectural complexity but also in terms of the types of mechanisms they can account for. For instance, while some of them try to precisely model neurophysiological mechanisms and thus make use of biologically plausible or inspired elements, other models' aim is to account for the possible underlying computations or to predict behavioral data without specifically taking into account biological constraints. For instance, while oscillatory-based models of speech parsing incorporate oscillatory mechanisms inspired from experimental observations,

deep neural networks based models such as the Transformers make use of backpropagation to adjust their weights during training phases. Whether this specific mechanism is biologically speaking realistic is still debated (Whittington & Bogacz, 2019). Moreover, such language models also make use of enormous amounts of text corpora to predict missing words, which is probably not how humans learn language.

From modelling a specific processing step in a specific behavioral task to end-to-end modelling which can be applied and generalized to diverses conditions (different tasks, languages, etc…), language models are a powerful tool. However, one needs to define the precise question to answer and the level of explanation to the question one wants to address in order to use models that suit one's needs.

3.3 Neurophysiological approaches

Speech is an inherently temporal signal. The neural mechanisms supporting language understanding must therefore proceed very quickly. This temporal aspect therefore limits the recording methodologies that can be used to track the underlying processes. Moreover, speech and language are specific to humans. They cannot be studied properly in animal models which constitute another constraint for researchers. Indeed, only a small part of all recording techniques used in Neuroscience are suited to human study. They include functional magnetic resonance imaging, electro- and magneto-encephalograpy. Whereas the former has a high spatial resolution well adapted to characterize language networks, the latter has a high temporal resolution better suited to record the successive brain computations supporting comprehension. Functional magnetic resonance imaging records the change in oxygen concentration within the blood flow across brain regions. This classical hemodynamic response is related to energy consumption by brain cells. Electro and magneto encephalograpy record electrical currents and magnetic fields that result from the simultaneous activity of a large number of neurons, respectively. As these recording modalities are non-invasive they are largely used in human studies. In rare cases, it is possible to record human brain activity directly from inside the brain. Electrocorticography enables recording electrical activity thanks to grids placed directly on the brain tissue. Stereo-electroencephalography (or intracranial electroencephalography) is characterised by the implantation of electrodes within the brain. These invasive methods are only used in clinical settings for particular conditions such as refractory epilepsy. I will mainly describe the last methodology in the rest of this part as the first experimental work in this thesis is based on data acquired with this modality in the epilepsy unit at La Timone hospital.

i. Stereo-electroencephalography

Stereo-electroencephalography was developed by Talairach and Bancaud and emerged as a completely new technique and method with a new philosophy in the mid-1960s in France (Talairach, Bancaud, Bonis, Szikla, & Tournoux, 1962). It involves the implantation of multi-contact depth electrodes inside the brain and allows the recording of coherent activity of local neuronal populations in the vicinity of the recording sites. The electrodes can be stereotactically implanted: they are placed in specific brain structures by reference to a standard atlas. Electrodes usually contain between 5 and 15 contacts that are 2 mm long and spaced from each other by 1.5 mm.

It is a type of presurgical investigation suitable for all types of drug-resistant epilepsies, used to record and map the epileptogenic and propagation networks. It enables the recording of spontaneous seizures, the recording of neural activity during sensory stimulation, and direct cortical electrical stimulation to trigger seizures (part of the localization process) and perform functional mapping in the same areas (to evaluate possible contraindications due to sensory or cognitive impairment such as language deficits). There is no "standard" electrode implantation. Implantation strategy depends only on the hypotheses made about the anatomical location of the epileptogenic zone, with the aim of defining subsequent cortectomy. Electrode positions are therefore not standardized across patients but chosen according to individual clinical characteristics. One of the most frequent regions to be explored is the perisylvian region in order to know for instance if the patient's epilepsy is temporal, temporo-perisylvian, or perisylvian. Perisylvian epilepsy can be localized in insular cortex, frontal opercular, parietal opercular, temporal opercular cortices and superior temporal and supramarginal gyri. In this region, electrodes are implanted orthogonally to the cortical surface in order to record along one single electrode both posterior insula, tip of Heschl's gyrus and planum temporale, for example. Another trajectory is to implant more anteriorly the superior temporal gyrus and the ventral insula. The perisylvian region is also implanted in the hypothesis of inferior parietal lobule, pericentral, or ventral prefrontal/premotor epilepsy. The transverse gyrus (Heschl's gyrus, which includes the auditory cortex) is critically important because of its connections with lower central regions and the inferior frontal gyrus: it is a pathway through which seizures from the temporal pole and the anterior superior temporal gyrus generalize. The implantation of electrodes is performed according to the Bancaud and Talairach stereotactic method (Talairach et al., 1962). The stereotaxic localization and the implantation of the intracerebral electrodes are done at the same time. All the anatomic information is merged in the stereotaxic space. Most of the electrodes are orthogonally implanted through the double talairach grid. The entry point of the electrodes is determined on the basis of the disposition of the vessels in the area. Finally, electrodes are implanted percutaneously by twist drill, and fixed using anchor bolts (Figure 15). Patients can be either implanted unilaterally or bilaterally.

Because of its invasiveness, intracranial recordings are applied exclusively for clinical purposes; nevertheless, the implantation time which in the unit lasts usually 14 days, provides a unique opportunity to investigate human brain functions and in particular language and

speech processing. As the temporal lobe is often implicated in refractory epilepsy resulting in a large electrode coverage of this area and as this brain region contains important hubs for speech comprehension, this is a rare and incredible window into language processing. The recorded signal is of high quality with a very high signal-to-noise ratio as it is located just next to the neural generators. Although the temporal and spatial resolution is very good, the spatial sampling is not homogenous and critically depends on clinical purposes. However, some of the most important contributions to the neural bases of language processing from the last 10 years capitalized on such an approach (Chang et al., 2010; Ding et al., 2016; Hamilton et al., 2018, 2021; Johnson, Kam, Tzovara, & Knight, 2020).

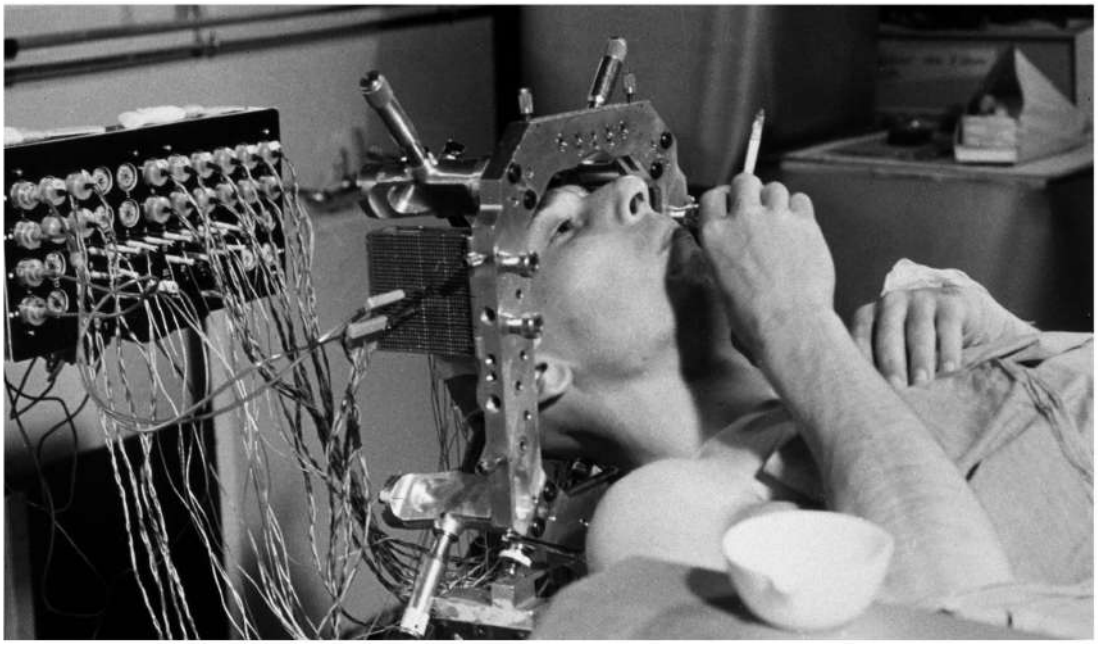


Figure 15. An epileptic patient undergoing surgery with stereo-electroencephalography. His head is maintained thanks to a stereotaxic frame and electrodes are inserted and positioned through it, based on the Talairach atlas. Archives of the Institut de Neurosciences des systèmes.

3.4 PhD Rationale

Understanding the fundamental mechanisms underlying speech comprehension has been a major goal of cognitive and neural sciences for more than 100 years. Multiple and complementary perspectives have been adopted in this regard.

From a cognitive viewpoint, the speech signal can be described at various levels of analysis as previously acknowledged in this dissertation. As a result, in the last decades, a large amount of experimental work has been conducted with different theoretical views and within various fields of investigation such as acoustics, linguistics and cognitive neuroscience for instance. Although a complete characterisation of speech critical aspects for comprehension is still missing, peculiar speech features have been proposed to be highly influential for understanding. Temporal speech envelope for instance has received lots of attention and seems to be very important for intelligibility. The spectrotemporal framework reviewed in the first chapter is also a promising path toward better understanding of the complexity of the speech signal. Another interesting avenue in the comprehension of speech and language is their probabilistic and highly organized structural nature.

On the other hand, and thanks to the development of (non-)invasive neuroimaging and recording methods, our characterisation of the neural substrates supporting speech and language processing has drastically increased in recent years. Departing from an historical highly modular and static vision based on documented pathological cases, the current models of language processing highlights the dynamical and distributive aspects of the complex computations at play during comprehension. Difficult questions still remain, one of which is concerned with the asymmetric processes taking place during speech processing. Another puzzle is related to the oscillatory nature of the brain response to spoken language and its link to understanding.

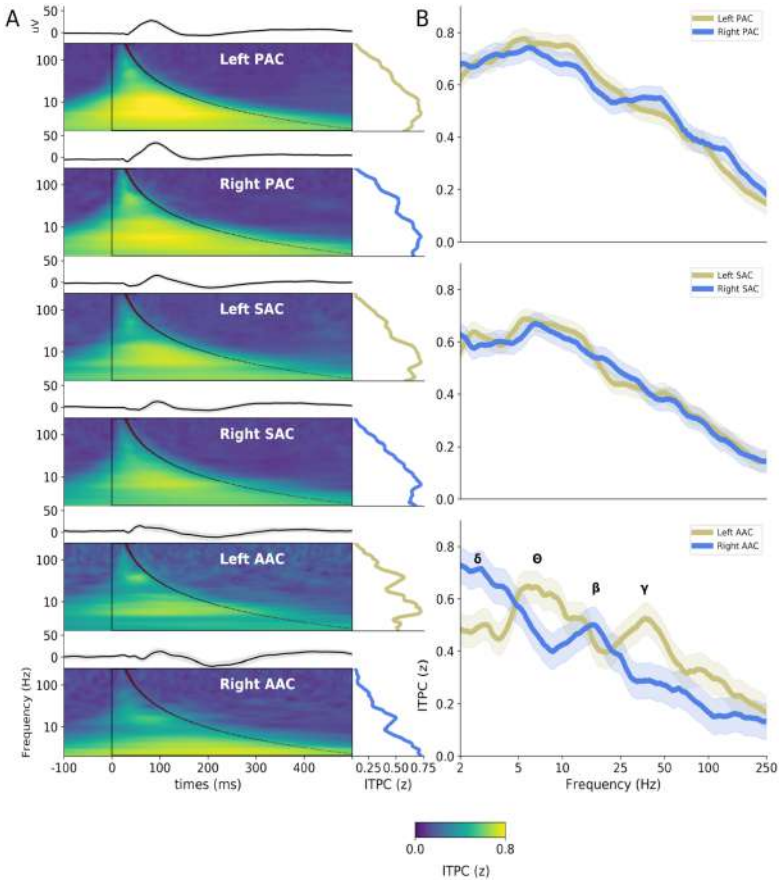
Determining both the neurophysiological and cognitive constraints at play during speech processing is an appealing approach for gaining critical insights into this highly complex cognitive function. In the present work, we try to address several limiting factors inherent to previous studies. First, using intracranial data from a large cohort of epileptic patients we aimed at offering a characterisation of the auditory cortical neural activity at a high temporal and spatial resolution. Such a detailed account of the auditory cortical hierarchy implications in auditory processing has been rendered difficult due to inherent limitations of non-invasive recording methods. Moreover, using different levels of auditory stimuli complexity (tones, syllables), we were able to record auditory areas' intrinsic constraints as well as characterizing the interaction between intrinsic and external activity induced by the auditory stimuli used. The objective of the second study was to determine the cognitive constraints implicated in speech comprehension. The main limitations of previous works in this endeavor include a lack of integrative framework and measurement scale to assess concurrently multiple constraints potentially limiting comprehension of spoken language. Capitalizing on the concept of channel capacity, the development of a highly sensitive behavioral paradigm and carefully designed spoken materials, we shed light on the multifactorial nature of the constraints limiting understanding. By taking complementary and

parallel approaches, the experimental contributions of this PhD refined our understanding of the capacity and limitations of our perceptual system to process speech and language.

Part II: Experimental contributions

Study 1

Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy



PLOS BIOLOGY

RESEARCH ARTICLE

Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy

Jérémy Giroud[1], Agnès Trébuchon[1,2], Daniele Schön[1], Patrick Marquis[1], Catherine Liegeois-Chauvel[1,3], David Poeppel[4,5], Benjamin Morillon[1]*

1 Aix Marseille University, Inserm, INS, Inst Neurosci Syst, Marseille, France, **2** APHM, Hôpital de la Timone, Service de Neurophysiologie Clinique, Marseille, France, **3** Cleveland Clinic Neurological Institute, Epilepsy Center, Cleveland, Ohio, United States of America, **4** Department of Neuroscience, Max-Planck-Institute for Empirical Aesthetics, Frankfurt am Main, Germany, **5** Department of Psychology and Center for Neural Science, New York University, New York, New York, United States of America

* bnmorillon@gmail.com

Citation: Giroud J, Trébuchon A, Schön D, Marquis P, Liegeois-Chauvel C, Poeppel D, et al. (2020) Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. PLoS Biol 18(3): e3000207. https://doi.org/10.1371/journal.pbio.3000207

Academic Editor: Timothy D. Griffiths, Newcastle University Medical School, UNITED KINGDOM

Received: March 7, 2019

Accepted: February 13, 2020

Published: March 2, 2020

Data Availability Statement: Numerical data supporting this study are available on GitHub: https://github.com/DCP-INS/asymmetric-sampling.

Funding: BM was supported by grants ANR-16-CONV-0002 (ILCB) and ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Speech perception is mediated by both left and right auditory cortices but with differential sensitivity to specific acoustic information contained in the speech signal. A detailed description of this functional asymmetry is missing, and the underlying models are widely debated. We analyzed cortical responses from 96 epilepsy patients with electrode implantation in left or right primary, secondary, and/or association auditory cortex (AAC). We presented short acoustic transients to noninvasively estimate the dynamical properties of multiple functional regions along the auditory cortical hierarchy. We show remarkably similar bimodal spectral response profiles in left and right primary and secondary regions, with evoked activity composed of dynamics in the theta (around 4–8 Hz) and beta–gamma (around 15–40 Hz) ranges. Beyond these first cortical levels of auditory processing, a hemispheric asymmetry emerged, with delta and beta band (3/15 Hz) responsivity prevailing in the right hemisphere and theta and gamma band (6/40 Hz) activity prevailing in the left. This asymmetry is also present during syllables presentation, but the evoked responses in AAC are more heterogeneous, with the co-occurrence of alpha (around 10 Hz) and gamma (>25 Hz) activity bilaterally. These intracranial data provide a more fine-grained and nuanced characterization of cortical auditory processing in the 2 hemispheres, shedding light on the neural dynamics that potentially shape auditory and speech processing at different levels of the cortical hierarchy.

Introduction

Contrary to the classic neuropsychological perspective, speech processing is now known to be distributed across the 2 hemispheres, with some models positing a leftward dominance for verbal comprehension and a rightward dominance for processing suprasegmental features, including aspects of prosody or voice processing [1]. The origin and function of lateralization continues to be vigorously debated, for example, with regard to its domain-general or domain-specific nature [2,3]. The former view predicts that lateralization of speech processing (and auditory processing, in general) originates in general-purpose mechanisms sensitive to the

Abbreviations: AAC, association auditory cortex; AEP, auditory evoked potential; AST, asymmetric sampling in time; ERP, event-related potential; ITPC, intertrial phase coherence; LOOCV, leave-one-out cross validation; MNI, Montreal Neurological Institute; MSE, mean squared error; NMF, non-negative matrix factorization; PAC, primary auditory cortex; ROI, region of interest; SAC, secondary auditory cortex; SEEG, stereotactic electroencephalography; STG, superior temporal gyrus.

low-level acoustic features present in speech. The domain-specific view postulates that speech is processed in a dedicated system lateralized to the left hemisphere. On this view, processing critically depends on the specific linguistic properties of a stimulus. Crucial to this debate is thus proper understanding of the distinctive sensitivity of the left and right auditory cortical regions to acoustic features, which should be grounded in characteristic anatomic-functional signatures.

There exists suggestive neuroanatomical evidence for structural differences between the left and right auditory cortex. The primary auditory cortex (A1, BA41) is larger in the left hemisphere, with a higher density of gray and white matter [4]. Moreover, the left auditory cortex contains larger cortical columns with a higher number of large pyramidal cells in cortical layer III than its right counterpart [5]. Those differences in cytoarchitectonic organization should coexist with electrophysiological and functional differences between auditory regions. Building on such observations, the asymmetric sampling in time (AST) hypothesis made several interrelated predictions related to the characteristics of auditory information processing at the cortical level [6]. The main tenets of the original AST hypothesis regarding anatomofunctional specifications can be synthetized as follows:

1. The human auditory system employs (at least) a two-timescale processing mode, characterized by oscillatory cycles that can be viewed as individual computational units. These 2 timescales operate in the low-gamma (around 25–50 Hz) and theta (around 4–8 Hz) frequency ranges, corresponding, respectively, to temporal integration windows of approximately 30 ms and 200 ms. Such temporal multiplexing allows the system to process in parallel acoustic information using 2 complementary algorithmic strategies, optimized to encode complementary spectrotemporal characteristic of sounds. This prediction—that sounds are processed at preferred and specific timescales—has received support from both auditory and speech-specific paradigms [7–15].

2. This dual-timescale processing operates in both hemispheres, but the ratio of neural ensembles dedicated to the processing of each timescale differs between left and right hemispheres. Indeed, while the left auditory cortex would preferentially process auditory streams using a short temporal integration window (30 ms), the right auditory cortex would preferentially sample information using a long temporal integration window (200 ms). Previous findings reported that left and right cortical auditory regions exhibit differences in their intrinsic oscillatory activity [16–18]. A relative leftward dominance of low-gamma neural oscillations and/or rightward dominance of theta oscillations is also visible during sensory stimulation [17,19,20]. This asymmetry is, moreover, reflected in the sensitivity of the left and right auditory cortex to different spectrotemporal modulations of sounds, with a leftward dominance for fast temporal modulations and/or a rightward dominance for slow temporal modulations [12,21–28].

3. The electrophysiological signature of this asymmetry emerges outside of primary auditory regions. The AST hypothesis in its original conception posited that at the level of core auditory cortex, there is no obvious functional asymmetry, but that beyond this first stage of cortical processing, a functional asymmetry should be visible, namely in the left and right association auditory regions. This last point has also received some empirical support [12,24,25,27].

While each of these predictions has received experimental support, they are also vigorously debated. In particular, one concern relates to the specificity of the left temporal lobe for faster temporal modulations. Some authors have suggested that most published results can be interpreted in an alternative framework, wherein only the right temporal lobe shows a marked

preference for certain properties of sounds (for example, longer durations or variations in pitch [3,29]). Moreover, in contrast with the AST hypothesis, some authors suggested that the hemispheric asymmetry may stem from core auditory areas (Heschl's gyrus) and not association cortex [16–18,21,22,30–32]. The conflicting results may be due to differences in paradigms and stimuli, as well as the resolution of the imaging instruments employed. However, anatomical and lesion studies also indicate hemispheric asymmetries at the level of the primary cortex [5,33]. Discrepancies may thus also arise from the fact that this asymmetry probably takes different forms along the auditory cortical pathway, with a more subtle functional signature at early cortical stages and more striking qualitative differences at later processing stages (see, for example, [17]). Finally—and this is a crucial aspect of the theory—the duration of these temporal integration windows was never precisely characterized physiologically with high-resolution data.

To more sensitively test the predictions of the AST hypothesis and overcome some of the difficulties in acquiring decisive data to characterize the signature of auditory hemispheric lateralization at both high spatial and temporal (hence also spectral) resolutions, we combined 2 innovative experimental approaches to noninvasively map the dynamical properties of distinct cortical auditory areas. 1) The brain is often described as a dynamical system oscillating at multiple frequencies [34,35]. Previous work has envisioned the event-related potential (ERP) as the impulse, stereotyped response of the brain. In such a framework, the dynamical properties of neuronal responses to external perturbations is assumed to depend both on the fined-grained structural constraints of the selective brain region investigated and on the characteristics of the stimulus [36,37]. Probing the brain with a short acoustic transient or impulse—i.e., a signal without any temporal dynamics—and examining the spectrotemporal properties of its evoked response therefore unveils the intrinsic dynamics of the auditory areas investigated. This method corresponds to a noninvasive mapping of the dynamical properties of specific cortical microcircuits [36]. 2) Thanks to the granularity offered by human intracranial recordings, a systematic investigation of the stereotyped response evoked by a brief (30-ms) pure tone can be performed in distinct cortical auditory areas. In addition to revealing the intrinsic dynamics of each region, it allows characterizing the functional asymmetry along the auditory pathway with a high spatial resolution. In the present study, we capitalize on data acquired from 96 epileptic patients, implanted for clinical evaluation at various stages of the auditory cortical hierarchy. Our results show the natural spectral profile of neural activity in left and right primary, secondary, and association cortical auditory regions, thus enabling a detailed characterization of the potential interhemispheric functional differences and dynamics at play during auditory processing.

Results

Data from 96 epileptic patients implanted with depth macroelectrodes located in left and right primary, secondary, and association auditory cortex (PAC, SAC, and AAC, respectively) were analyzed (Fig 1) [38–41]. Auditory areas were defined with a functional localizer [30,38,39,42]. PAC, SAC, and AAC, respectively, correspond to the posteromedial portion of Heschl's gyrus (A1, medial belt and lateral belt areas; BA41, anterior to Heschl's sulcus), the lateral posterior superior temporal gyrus (STG; parabelt area; anterior portion of BA42, posterior to Heschl's sulcus), and the lateral anterior STG (area A4; anterior BA22) [40,43]. 78% of the patients had a typical language lateralization in the left hemisphere (see Materials and methods). Left hemisphere dominance for language is usually observed in approximately 90% of healthy individuals and in 70% of epileptic patients [44]. Patients participated in a perceptual experiment during which they passively listened to pure tones and syllables (see Materials and methods).

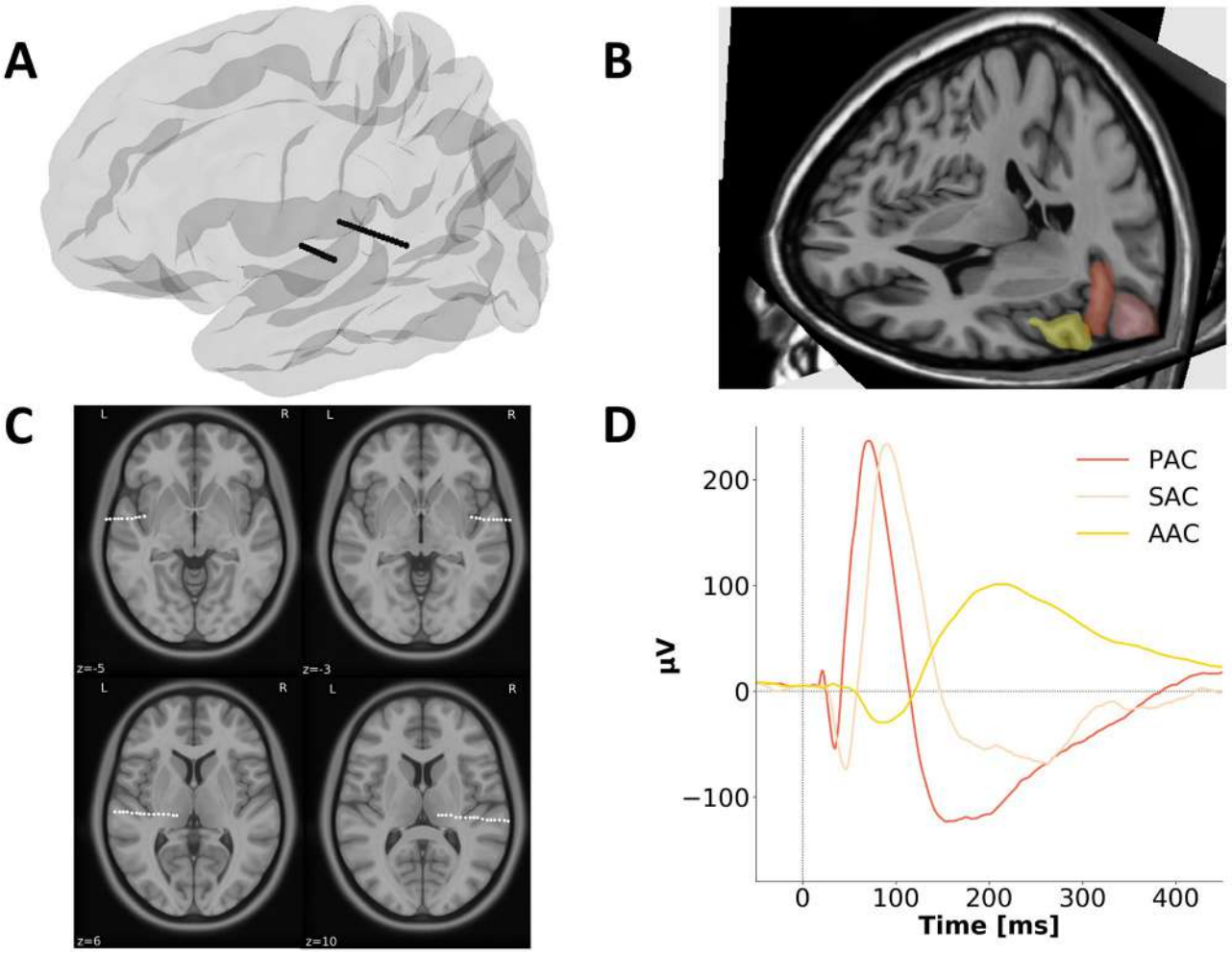


Fig 1. Example of electrodes and contact positions with characteristic AEPs in exemplar patients. (A) Schematic example of typical SEEG recording electrodes shown in a 3D view of the left hemisphere on a template brain in an MNI space. The posterior electrode is composed of 15 contacts and targets core auditory regions, while the anterior electrode, composed of 5–15 contacts, targets association auditory regions. (B) MRI scan showing the location of the 3 ROIs in the left hemisphere. Red: PAC, the posteromedial portion of Heschl's gyrus (A1, medial belt and lateral belt areas; see [43]); pink: SAC, the lateral posterior STG (parabelt area); yellow: AAC, the lateral anterior STG (area A4). (C) Representative examples of SEEG recording electrodes shown in an axial view of a template brain in an MNI space, targeting left or right association (upper panels) or core (lower panels) auditory regions. (D) AEPs in response to pure tones from a representative patient for each ROI. The axes are color-coded according to the locations displayed in B. Electrode contacts used along the shaft were selected based on their anatomical location and functional responses (typical shape and latencies of evoked responses; see Materials and methods). AAC, association auditory cortex; AEP, auditory evoked potential; MNI, Montreal Neurological Institute; PAC, primary auditory cortex; ROI, region of interest; SAC, secondary auditory cortex; SEEG, stereotactic electroencephalography; STG, superior temporal gyrus.

https://doi.org/10.1371/journal.pbio.3000207.g001

Spectral characteristics of the evoked response to transient pure tones

To investigate the fine-grained temporal constraints of the first cortical stages of the auditory processing hierarchy, we first analyzed the evoked, stereotyped (i.e., identical across trials) responses to transient acoustic impulses (30-ms duration pure tones, presented at 0.5 or 1 kHz). A time-frequency representation of the evoked responses, as computed through inter-trial phase coherence (ITPC; Fig 2A), demonstrates the presence of a dynamical response composed of multiple spectral modes (i.e., time constants), which, moreover, could differ between regions of interest (ROIs). These responses were limited in time and homogenous, and their

spectral profile was best captured by averaging ITPC values over time (see Materials and methods).

In PAC and SAC, a group-level analysis (computed after normalizing individual ITPC spectra to minimize the potential impact of a few outliers on group results) revealed the presence of a simple evoked response profile, characterized by a main spectral maximum within the theta range (around 4–8 Hz; corresponding to a time constant of approximately 150 ms; Fig 2B). Importantly, this response profile was similar across left and right hemispheres. Conversely, a more complex pattern of response was visible in AAC, with the presence of 2 distinct salient spectral maxima that moreover differed between left and right hemispheres. Prominent ITPC peaks in the theta (4–8 Hz) and low-gamma (25–50 Hz) frequency ranges were visible in left AAC; the right counterpart was characterized by peaks in the delta (1–4 Hz) and beta (13–30 Hz) frequency ranges (Fig 2B).

To better characterize the time constants of the neural processes occurring at each putative step of the auditory cortical hierarchy, we extracted for each patient and ROI the 2 highest local maxima of the ITPC spectrum (between 2–250 Hz; Fig 3 and S1 Fig). This analysis substantiates the finding that in PAC and SAC, the evoked response was dominated by an ITPC peak in the theta range (around 4–8 Hz) and highlights that a secondary peak emerged in the beta/gamma range (around 15–40 Hz; see interindividual spectral distributions in Fig 3). At these earlier cortical stages, the frequency of the 2 main ITPC peaks did not differ significantly across hemispheres, neither in PAC (Mann–Whitney U test: first peak, $U = 485.0$, $p = 0.29$; second peak, $U = 496.5$, $p = 0.35$) nor in SAC (first peak, $U = 484.0$, $p = 0.12$; second peak, $U = 502.0$, $p = 0.17$). In contrast, in AAC, a more complex and significantly asymmetric response profile emerged. We confirmed that the evoked response was characterized by higher-frequency peaks in left than right AAC, with respectively theta/gamma (around 8/35 Hz) peaks in left and delta/beta (around 4/15 Hz) peaks in right AAC. Interhemispheric comparison of the frequency of the 2 main ITPC peaks confirmed that this asymmetry was significant (first peak, $U = 34.5$, $p < 0.05$; second peak, $U = 36.5$, $p < 0.05$). To confirm the robustness of these findings, we reanalyzed the evoked response to pure tones separately for 0.5 and 1 kHz pure tones and observed the exact same ITPC spectral profile for each ROI, independently of the frequency of the pure tone (S2 Fig).

Interindividual consistency of the spectral response profile in left and right AAC

While our previous analyses were focusing on the frequency of the 2 main ITPC peaks, we next investigated hemispheric asymmetry while taking advantage of the entire ITPC spectrum (between 2–250 Hz). A leave-one-out cross-validation (LOOCV) procedure was used to assess the similarity of each individual ITPC spectrum with the left and right ITPC patterns obtained at the group level (Fig 4A and 4B). The rationale of this analysis is that group-level ITPC patterns are a good approximate of a prototypical response and can thus be used as "models" upon which individual data can be compared. Briefly, for each individual ITPC spectrum, we estimated its mean squared error (MSE, i.e., the error of fit) relative to both left and right AAC models (i.e., group-level ITPC patterns; see Materials and methods). On average, left AAC ITPC spectra were more similar to the left AAC model than the right one (unpaired Mann–Whitney U test: $U = 36.0$, $p = 0.020$). On the contrary, right AAC ITPC spectra were more similar to the right than the left AAC model ($U = 40.0$, $p = 0.034$). Moreover, the only patient implanted bilaterally in AAC showed an asymmetric response profile compatible with the group-level ITPC spectra (Fig 4C).

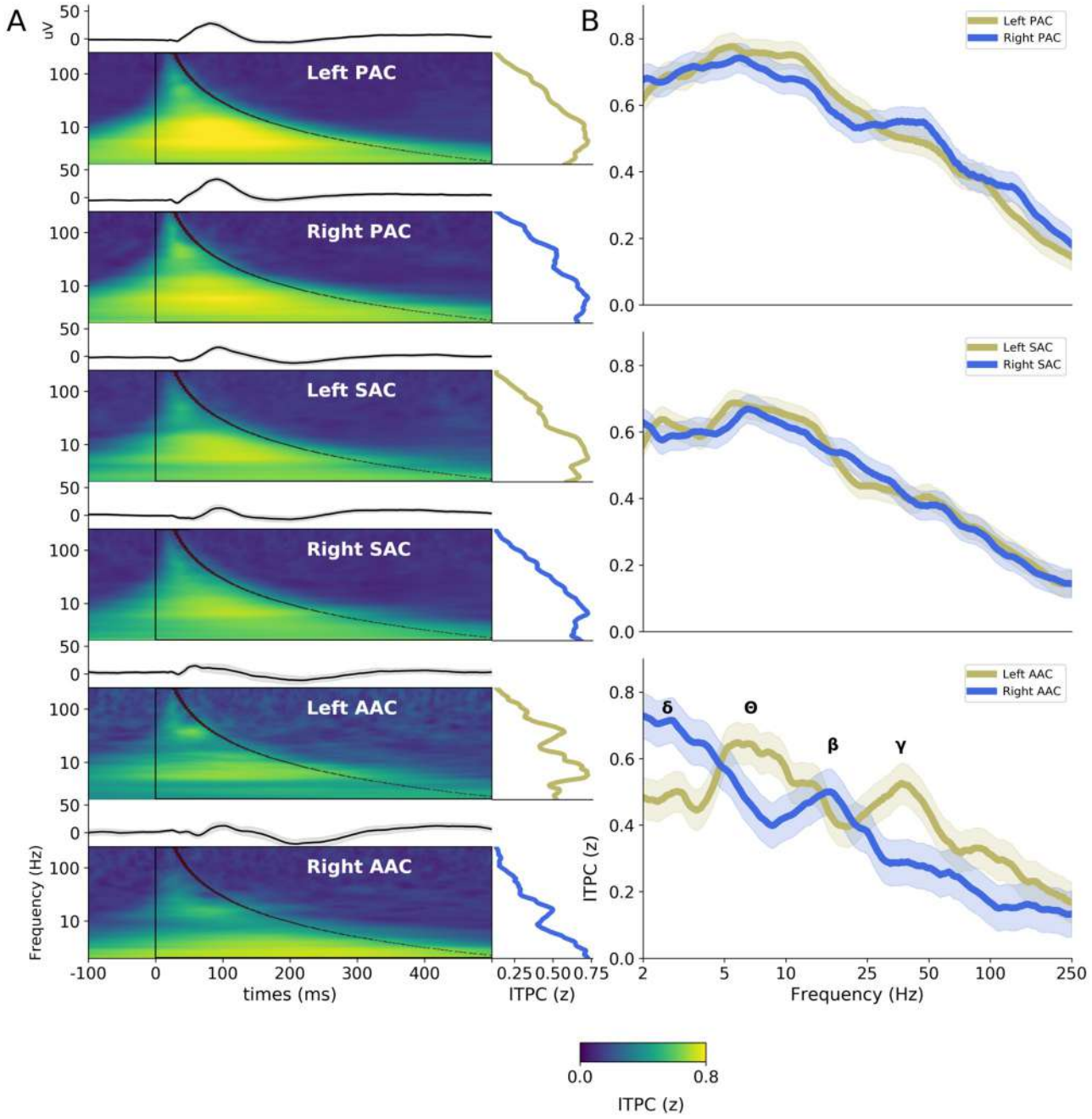


Fig 2. Evoked activity in response to pure tones (0.5 kHz and 1 kHz merged) in hierarchically organized auditory areas. (A) AEPs (top panels) and ITPC (lower panels) in response to pure tones, averaged across patients, in the 6 auditory ROIs (PAC, SAC, and AAC, in left and right hemispheres). Right insets indicate the spectrum of the ITPC averaged over one oscillatory cycle (black overlay). (B) Interhemispheric comparison of the ITPC spectra in PAC, SAC, and AAC. Shaded areas indicate SEM. Greek letters indicate the main peaks observed in AAC (δ: delta 1–4 Hz; θ: theta 4–8 Hz; β: beta 14–30 Hz; γ: low-gamma 25–45 Hz). Number of patients recorded at each location: left PAC = 39, right PAC = 27; left SAC = 40, right SAC = 29; left AAC = 12, right AAC = 12. AAC, association auditory cortex; AEP, auditory evoked potential; ITPC, intertrial phase coherence; PAC, primary auditory cortex; ROI, region of interest; SAC, secondary auditory cortex.

https://doi.org/10.1371/journal.pbio.3000207.g002

Then, we investigated whether this asymmetric response in AAC, visible at the group level, was robust at the individual level (Fig 4D). This methodology aims at determining whether the implantation hemisphere (left or right) of a patient can be predicted from the functional response of its AAC region to a brief auditory stimulation. This analysis revealed that for 16

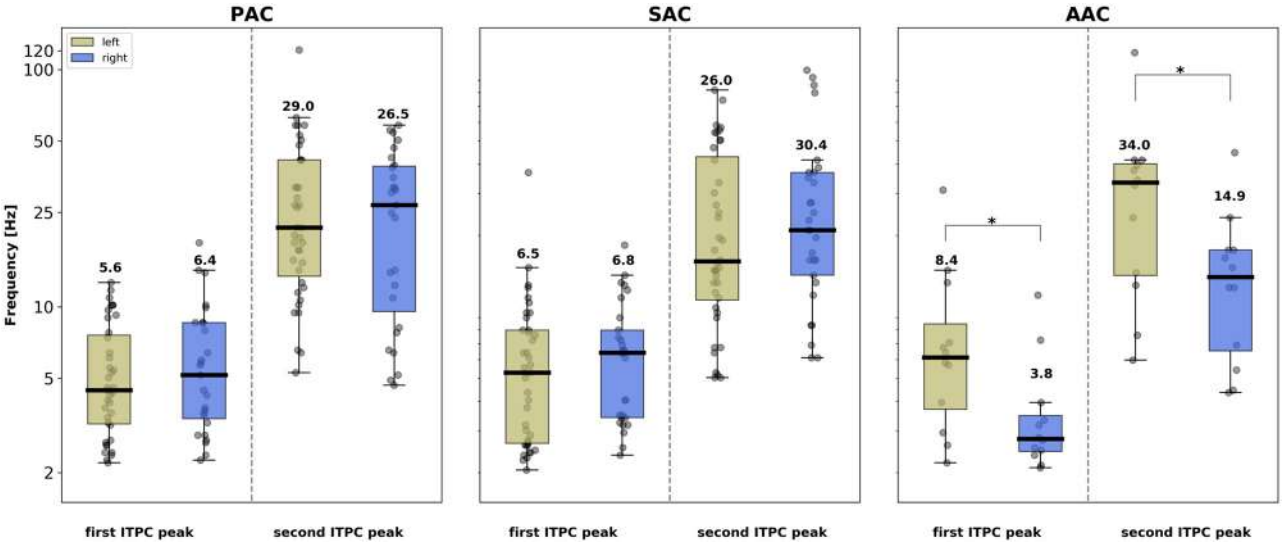


Fig 3. Interhemispheric comparison of the frequency of the 2 main ITPC peaks from the evoked response to pure tones (0.5 kHz and 1 kHz merged) in the different auditory areas. Individual peaks were identified as the 2 highest noncontiguous maxima of the ITPC spectrum (between 2–250 Hz). Frequency (in Hz) is denoted on the y-axis. Box plots: contour lines reflect the first and third quartile of the distribution, and the thick central line denotes the median. Numbers indicate the mean. Gray dots correspond to individual data. Stars indicate significant interhemispheric differences (unpaired Mann–Whitney U Tests, $p < 0.05$). AAC, association auditory cortex; ITPC, intertrial phase coherence; PAC, primary auditory cortex; SAC, secondary auditory cortex.

https://doi.org/10.1371/journal.pbio.3000207.g003

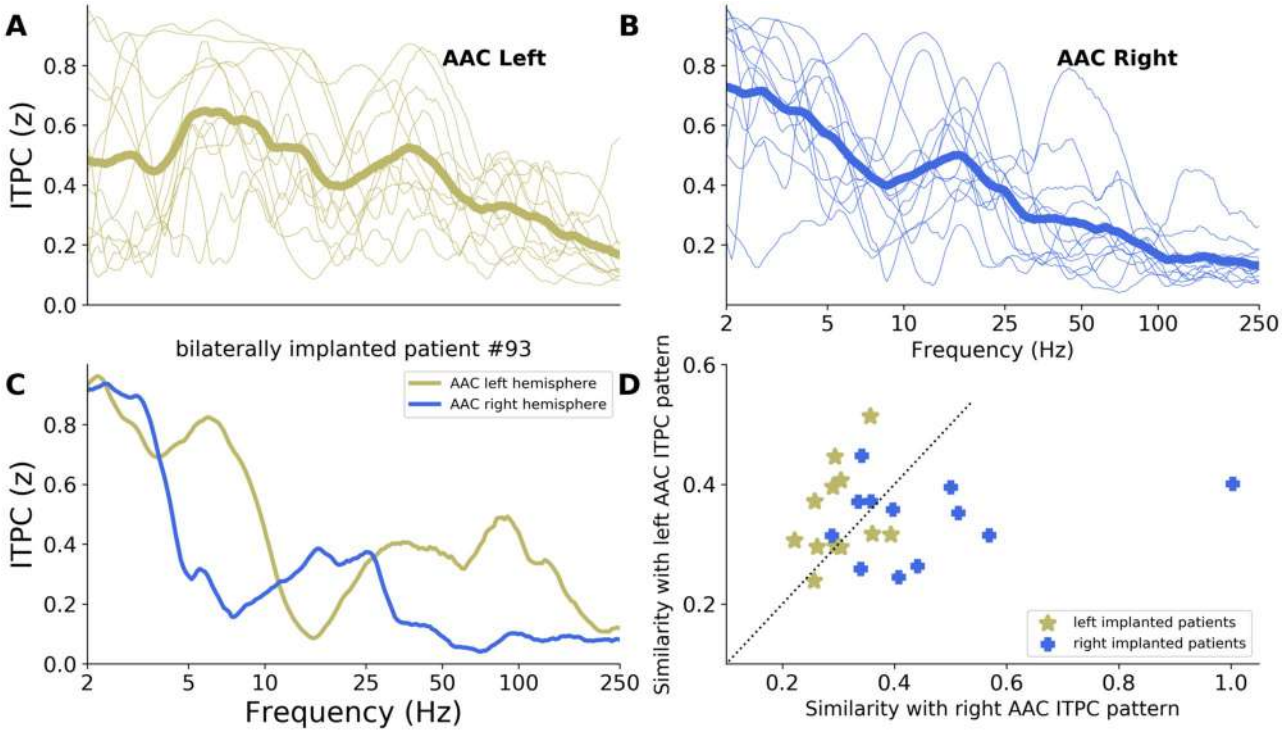


Fig 4. Interindividual consistency of the ITPC spectra in response to pure tones in left and right AAC. (A–B) Individual ITPC spectra in response to pure tones (0.5 kHz and 1 kHz merged). Thick lines: group-level spectral pattern. Thin lines: individual data (A: left AAC, $n = 12$; B: right AAC, $n = 12$). (C) ITPC spectra of a patient implanted bilaterally in AAC. (D) Similarity of each individual ITPC spectrum with the group-level ITPC patterns obtained in left (y-axis) and right (x-axis) AAC (see Materials and methods). The dashed diagonal indicates equal similarity to both group-level ITPC patterns. AAC, association auditory cortex; ITPC, intertrial phase coherence.

https://doi.org/10.1371/journal.pbio.3000207.g004

out of the 24 (67%) patients implanted in AAC, their ITPC spectrum was more similar to the model (i.e., group-level ITPC pattern) of their hemisphere of implantation than to the one of the opposite hemisphere.

Amongst the 24 patients implanted in AAC, 4 had a nontypical language lateralization that was moreover complex, reflecting a bilateral organization of language functions. In a following analysis, we thus excluded them and recomputed the similarity analysis for the remaining 20 patients with a typical language lateralization in the left hemisphere. We observed that 14/20 (70%) patients had a spectral profile of response congruent with the ITPC pattern of their hemisphere of implantation (7/10 in left AAC, 7/10 in right AAC).

Prototypical ITPC spectral components of the AAC response

Next, we analyzed the entire set of electrode contacts implanted in AAC. On the one hand, this approach allows evaluating the potential impact of our selection criteria (restricted to the electrode contact with the largest auditory evoked potential [AEP] per patient) on the previous results. On the other hand, having a large number of recordings is well-suited to perform data- driven analyses. All contacts from all electrodes implanted in left and right AAC were regrouped and analyzed as previously to extract their ITPC spectra in response to pure tones. A non-negative matrix factorization (NMF) was then conducted on this extended data set. This unsupervised clustering method allows extracting the prototypical ITPC spectral components constituting this data set, in which left and right AAC were combined (see Materials and methods).

This analysis yielded 4 main ITPC spectral components, which altogether explained 63% of the variance of the data set. These components could be regrouped according to their spectral profile. Two components (#1 and #2) had main peaks in the delta (around 2.5 Hz) and/or beta (around 16 Hz) bands, and the other 2 components (#3 and #4) had main peaks in the theta (around 5 Hz) and gamma (around 45 Hz) bands (Fig 5A). This indicates that across the entire set of electrode contacts present in (left and right) AAC, the evoked responses to a brief acoustic stimulation are principally composed of delta/beta and theta/gamma bimodal spectral patterns.

We evaluated the relative weight of these components on left or right AAC responses (Fig 5B). Across electrode contacts, we observed a significant components by hemisphere interaction (unpaired t test: $t = 4.3$, $p < 0.001$), with components #3 + 4 (theta/gamma pattern) being significantly more present in left than right AAC ($t = 3.6$, $p < 0.001$), and components #1 + 2 (delta/beta pattern) being significantly more present in right than left AAC ($t = 3.2$, $p = 0.002$). To estimate whether this interaction was robust at the individual level, we extracted the contribution (relative weight) of each component to the individual data by averaging intrapatient electrode contacts (Fig 5C). We observed that the responses of 8/12 (67%) patients implanted in the left hemisphere were predominantly composed of the theta/gamma components (#3 and 4), while the responses of 10/12 (83%) patients implanted in the right hemisphere were predominantly composed of the delta/beta components (#1 and 2).

We finally replicated this analysis on the 20 patients implanted in AAC that had a typical language lateralization in the left hemisphere. We observed that 16/20 (80%) patients had a response profile predominantly composed of the expected spectral components (7/10 dominated by theta/gamma components [#3 and 4] in left AAC, 9/10 dominated by delta/beta components [#1 and 2] in right AAC). This new analysis (clustering of entire set of electrode contacts) thus reveals that the specificity of the spectral responses in left and right AAC is actually quite robust at the individual level. Overall, this confirms the existence of a functional

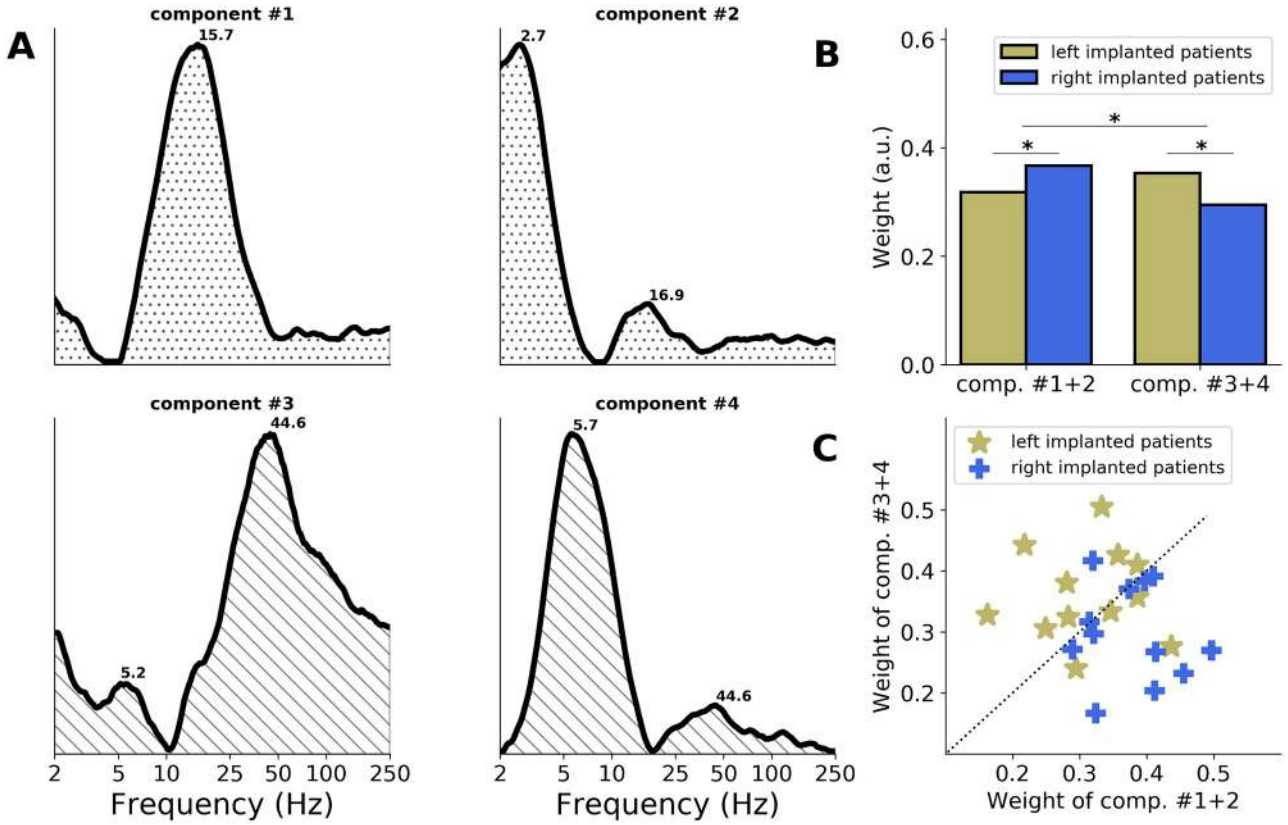


Fig 5. Main spectral components of the AAC response to pure tones. The NMF method applied to the ITPC spectra in response to pure tones (0.5 kHz and 1 kHz merged) of the entire set of electrode contacts implanted in AAC (left and right combined; $n = 24$). (A) Four main ITPC spectral components estimated with the NMF (see Materials and methods). Numbers indicate the main frequency peaks of each component. Components can be regrouped according to their spectral profile, into "delta/beta" (around 2.5/16 Hz; components #1 and 2) and "theta/gamma" (around 5/45 Hz; components #3 and 4) patterns. (B) Respective contribution (relative weight) of components #1 and 2 and #3 and 4 to left and right AAC data. Stars indicate significant differences (unpaired *t* tests, $p < 0.05$). (C) Respective contribution (relative weight) of components #1 and 2 (x-axis) and #3 and 4 (y-axis) to individual data. Patients are sorted according to the hemisphere of implantation (yellow: left; blue: right). The dashed diagonal indicates equal weights. AAC, association auditory cortex; ITPC, intertrial phase coherence; NMF, non-negative matrix factorization.

https://doi.org/10.1371/journal.pbio.3000207.g005

asymmetry in AAC, with a predominant theta/gamma (around 5/45 Hz) activity in left AAC and a predominant delta/beta (around 2.5/16 Hz) activity in right AAC.

Interaction between stimulus and neural dynamics

To evaluate the interaction between stimulus and neural dynamics, the same analysis was carried on data recorded on the same patients during presentation of syllables (French /ba/ and /pa/; Fig 6). Importantly, these stimuli are characterized by more complex spectrotemporal dynamics than transient pure tones and carry linguistic information. Accordingly, we observed that the ITPC spectral response profile differed between pure tones and /ba/ and /pa/ stimuli, as predicted. Responses to syllables yielded less prominent and specific spectral peaks in the different ROIs, even in the latter stages of auditory processing (AAC). The maximum neural activity in response to syllable presentation was in the low frequency range (<20 Hz) and did not change across ROIs.

We extracted for each patient, ROI, and syllable type (/ba/ or /pa/) the 2 highest local maxima of the ITPC spectrum. Overall, the peak frequencies were similar across ROIs and syllable types, with ITPC spectra dominated by a main peak in the delta/theta range (around 4 Hz),

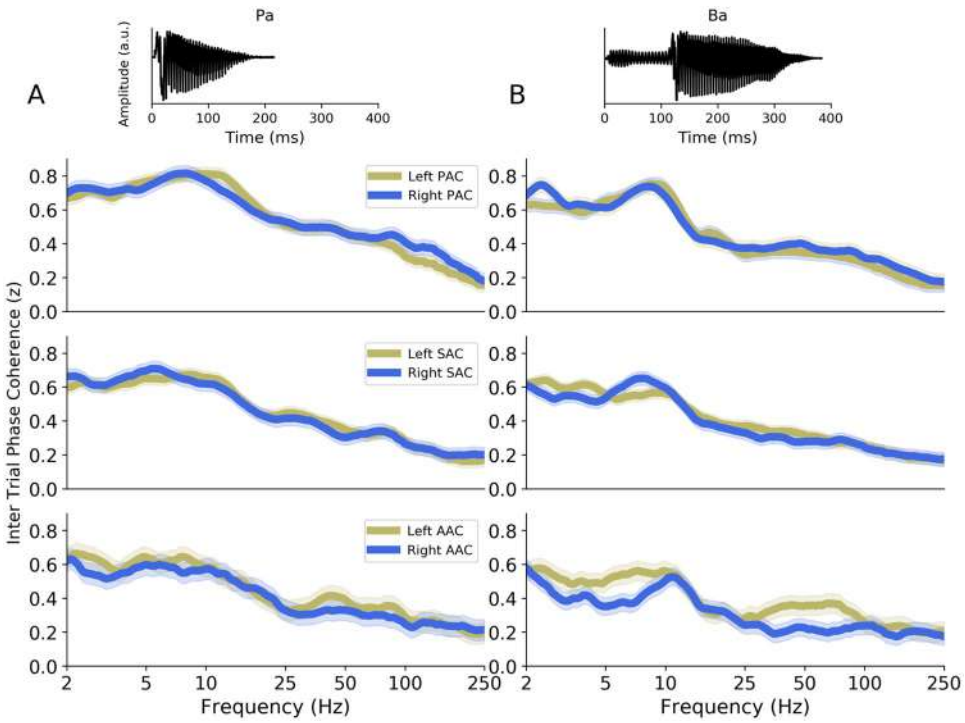


Fig 6. Evoked activity in response to French syllables /pa/ and /ba/ in the different auditory areas. (A–B) Top panels: acoustic waveform of the syllables. Lower panels: interhemispheric comparison of the ITPC spectra in response to the syllables (A) /pa/ and (B) /ba/, in PAC, SAC, and AAC. AAC, association auditory cortex; a.u., arbitrary unit; ITPC, intertrial phase coherence; PAC, primary auditory cortex; SAC, secondary auditory cortex.

https://doi.org/10.1371/journal.pbio.3000207.g006

with a secondary peak in the beta range (around 18 Hz). Crucially, the frequency of the 2 main ITPC peaks did not differ significantly across hemispheres in PAC, SAC, or AAC (unpaired Mann–Whitney U tests: all p-values > 0.09).

To better understand this result, we analyzed the entire set of electrode contacts implanted in AAC and performed an NMF clustering approach (as previously described). We extracted the 4 main ITPC spectral components, which altogether explained 55% of the variance of the data set. These components were characterized by 4 different spectral profiles (Fig 7A). While components #2 and #4 were, respectively, characterized by a "delta/beta" (around 2/13 Hz) and "theta/gamma" (around 5/48 Hz) pattern, components #1 and #3 had, respectively, an alpha (around 10 Hz) and gamma (>25 Hz) spectral profile, which were not observed in the responses to pure tones. These latter components can thus be interpreted as specific to syllables (i.e., more acoustically complex and/or linguistic) stimuli processing.

We evaluated the relative weight of these components on left or right AAC responses (Fig 7B). Across electrode contacts, we observed that 2 out of the 4 components were asymmetrically distributed. While components #1 (alpha) and #3 (gamma) were not significantly lateralized (unpaired t tests: component #1: $t = 0.7$, $p = 0.49$; component #3: $t = 1.5$, $p = 0.14$), component #4 (theta/gamma pattern) was significantly more present in left than right AAC ($t = 3.8$, $p < 0.001$), and component #2 (delta/beta pattern) was significantly more present in right than left AAC ($t = 4.0$, $p < 0.001$). We finally estimated whether this interaction was robust at the individual level by extracting the contribution (relative weight) of components #2 and #4 to the individual data (Fig 7C). We observed that only 54% of the patients implanted in AAC had a response profile predominantly composed of the expected spectral component (7/

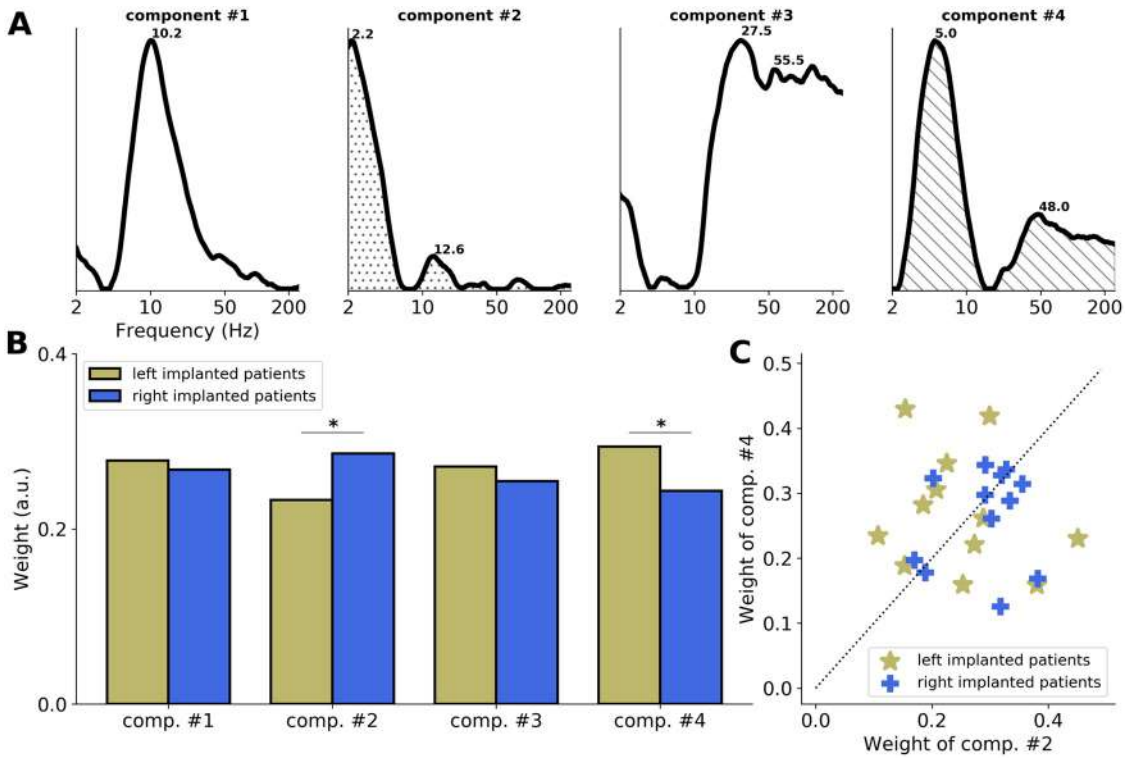


Fig 7. Main spectral components of the AAC response to syllables. The NMF method applied to the ITPC spectra in response to syllables (French /ba/ and /pa/ merged) of the entire set of electrode contacts implanted in AAC (left and right combined; $n = 24$). (A) Four main ITPC spectral components estimated with the NMF. Numbers indicate the main frequency peaks of each component. Components #2 and #4 are, respectively, characterized by a "delta/beta" (around 2/13 Hz) and "theta/gamma" (around 5/48 Hz) pattern, while components #1 and #3 have a spectral profile specific to syllable presentation. (B) Respective contribution (relative weight) of components #2 and #4 to left and right AAC data. Stars indicate significant differences (unpaired *t* tests, $p < 0.05$). (C) Respective contribution of components #2 (x-axis) and #4 (y-axis) to individual data. Patients are sorted according to the hemisphere of implantation (yellow: left; blue: right). The dashed diagonal indicates equal weights. AAC, association auditory cortex; a.u., arbitrary unit; ITPC, intertrial phase coherence; NMF, non-negative matrix factorization.

https://doi.org/10.1371/journal.pbio.3000207.g007

12 dominated by a theta/gamma component [#4] in left AAC; 6/12 dominated by a delta/beta component [#2] in right AAC).

This set of analyses first reveals that during perception of more complex stimuli—which carry a spectrotemporal dynamics and/or linguistic information—the evoked neural response is more heterogeneous, in particular with the emergence of alpha and gamma activity bilaterally in AAC (Fig 7). Second, it shows that the functionally asymmetric response observed during pure tone processing is also present during syllable processing, as revealed by the clustering analysis, with predominant theta/gamma (around 5/48 Hz) activity in left AAC and predominant delta/beta (around 2/13 Hz) activity in right AAC. However, this asymmetry is less salient, as evidenced by the absence of visible differences in the raw ITPC spectral profiles (Fig 6) and the difficulty to observe it at the individual level (Fig 7C).

Discussion

Derived from intracranial recordings from 96 epileptic patients and using transient auditory stimulation with tones and syllables, this study aimed to characterize the intrinsic timescales of auditory information processing at 3 stages of the auditory cortical hierarchy. The spatial and temporal precision offered by stereotactic electroencephalography (SEEG) recording enables a

meticulous description of the neural dynamics at each anatomic stage. Using transient acoustic stimulation allowed us to probe stereotyped evoked neural responses that uncover the intrinsic dynamics of the recorded areas. Using syllables, which are characterized by more complex spectrotemporal dynamics, allowed us to further investigate the interaction between stimulus and neural dynamics.

Our results reveal first, that the early cortical stages of auditory processing (PAC, SAC), when acoustically stimulated with transient pure tone stimuli, show characteristic bimodal spectral profiles. These functional responses were characterized by a main spectral peak in the theta range (around 4–8 Hz), with a secondary peak in the beta/gamma range (around 15–40 Hz; Fig 3). This finding, obtained with high-resolution data on a large cohort of patients, is consistent with previous findings obtained with intracranial recordings on a single case [18]. Moreover, the presence of 2 concomitant time constants in the dynamics of the evoked response is a strong evidence in favor of the AST framework, in which the auditory system makes use of a two-timescale processing mode to perceptually sample acoustic dynamics [9,13]. This specific bimodal neural signature also corroborates physiological descriptions of a natural interaction between low and high oscillatory frequencies (phase-amplitude coupling) at rest and during auditory stimulation at the level of local neural ensembles [45]. No hemispheric difference is apparent in core auditory regions, as evidenced by the similar bimodal spectral pattern elicited in left and right PAC and SAC. This is in contradistinction with previous findings describing functional asymmetries in core auditory areas [16–18,21,22,30–32]. One limitation of SEEG recording is the absence of whole-brain coverage. In particular, our functional characterization of early cortical auditory processing was limited to the posteromedial portion of Heschl's gyrus (PAC) and the lateral posterior STG (SAC; Fig 1). Hence, we cannot exclude the presence of a functional asymmetry in core auditory regions, notably in the lateral portion of Heschl's gyrus and the planum polare [46], which we did not sample. However, our results are compatible with current models of the functional organization of the core auditory cortex, which report relatively weak functional hemispheric differences [47].

Second, our results show the emergence of a strong functional asymmetry at the level of AAC. Using stimulation with short acoustic transients, we observed predominant theta/gamma (around 5/45 Hz) activity in left AAC and predominant delta/beta (around 2.5/16 Hz) activity in right AAC (Figs 2–5). Importantly, the clustering approach revealed that in AAC, the evoked responses to a brief acoustic stimulation are principally composed of these theta/gamma and delta/beta bimodal spectral patterns (Fig 5). Given the limitations/constraints imposed by the nature of the population under study, the emergence of a functional asymmetry in AAC was mostly demonstrated across patients at the group level. We, however, show that for 80% (16/20) of the patients implanted in AAC that had a typical language lateralization in the left hemisphere, the response profile was predominantly composed of the expected spectral components (theta/gamma in left AAC and delta/beta in right AAC). Moreover, our data set included a patient implanted bilaterally in AAC for whom we observed similar response profiles (Fig 4C), overall highlighting the robustness of the results at the individual level. Of note, because only 4 out of 24 patients implanted in AAC had a nontypical language lateralization that was also complex, reflecting a bilateral organization of language functions, it is difficult to estimate whether the observed functional asymmetry supports the lateralization of language functions.

Third, our results indicate that whereas a striking functional asymmetry in the higher auditory regions (AAC) is visible during brief acoustic stimulation, it is somehow obscured but nonetheless still present during the processing of more acoustically complex stimuli such as syllables. Indeed, the methodology used in the present study necessitates probing the brain with a short acoustic transient or impulse—i.e., a signal without any temporal dynamics.

However, the functional asymmetry observed during pure tone processing in AAC is also present during syllable processing, as revealed by the clustering analysis (Fig 7B). This asymmetry is less salient, as evidenced by the absence of visible differences in the raw ITPC spectral profiles (Fig 6) and the difficulty to observe it at the individual level (Fig 7C). One reason is that the evoked response to syllables is not only composed of theta/gamma and delta/beta bimodal spectral patterns but is also more heterogeneous, with the emergence of alpha and gamma activity (Fig 7A). Those complementary spectral modes may emerge from the elaborate interaction between the spectrotemporal acoustic features of the stimulus and the intrinsic neural activity and/or reflect the processing of linguistic (phonemic) information.

Most of our results—obtained with short acoustic transients and, to a lesser extent, with syllables—are in accordance with the main tenets of the original AST hypothesis, notably with point 1 (the observation of a bimodal spectral profile in all the areas investigated) and point 3 (the emergence of a functional asymmetry in association areas). However, they also reveal that this functional asymmetry does not simply correspond to a differential ratio of neural ensembles oscillating at theta and gamma rates (point 2) [6] but, in fact, corresponds to the involvement of distinct dynamics (theta/gamma versus delta/beta) in left and right hemispheres. Reframing the AST hypothesis after our findings could result in the following set of hypotheses:

1. The spectral profile of neural response in left AAC (but also bilateral PAC and SAC) could be linked to a recent model of coupled oscillators describing the sensory analysis of speech, in which low and high frequency oscillations operate in the theta and gamma ranges, respectively, and process in parallel acoustic information [48]. In this model, the tracking of slow speech fluctuations by theta oscillations and its coupling to gamma activity both appear as critical features for accurate speech encoding, underscoring the importance of a two-timescale processing mode for efficiently analyzing speech. Moreover, this model suggests that during speech perception, syllabic- and phonemic-scale computations operate in combination at a local cortical level of processing, which could correspond to the left AAC.

2. On the other hand, the presence of neural activity in the delta and beta ranges in right AAC is more puzzling. Previous studies claimed that parsing at the syllabic scale occurs bilaterally [49] or is even rightward lateralized [14,23]. However, the right auditory cortex is more sensitive to spectral than temporal modulations [27,28,50], and perception of prosody is a right-lateralized process [51]. Thus, our observation of a specific response dynamics in right AAC could be linked to neural mechanisms dedicated to the parsing of spectral acoustic dynamics. Prosodic phenomena at the level of intonation contours are an example of such a phenomenon; the successful perceptual analysis of spoken language requires the processing of the rhythmic and melodic variations in speech to gain knowledge about speaker's emotions and intentions [51]. The delta intrinsic timescale observed in right AAC would be particularly well-suited to the segmentation of prosodic cues because they naturally unfold at 0.5–3 Hz, as also argued in a recent computational model [52].

Overall, our results shed light on the neurofunctional architecture of cortical auditory processing and in particular on the specific processing timescales of different cortical areas. These general mechanisms are thought to apply to general auditory as well as speech perception. By integrating our findings to the AST hypothesis, we would speculate that syllabic and phonemic information is segmented in parallel locally, through coupled theta and gamma oscillations, while right-lateralized processes such as intonation contour or prosody perception would be segmented by delta (and beta) oscillations. The methodology we employed here is only suited to transient stimuli because longer stimuli with a spectrotemporal dynamics impose strong

temporal constraints on the neural activity, resulting in more heterogeneous and elaborate response profiles. It is thus an "intermediate" method between resting state and speech paradigms, allowing a more precise description of the natural dynamics at play throughout the auditory pathway.

Materials and methods

Ethics statement

The study was approved in accordance with the Declaration of Helsinki by the Institutional Review board of the French Institute of Health (IRB00003888). Patients provided written informed consent prior to the experimental session. Participation was voluntary, and none of these patients participated in a clinical trial.

Participants

96 patients (46 females) with pharmacoresistant epilepsy took part in the study. They were implanted with depth electrodes for clinical purpose at the Hôpital de la Timone (Marseille). Their native language was French. Neuropsychological assessments carried out before SEEG recordings indicated that all patients had intact language functions and met the criteria for normal hearing. None of them had their epileptogenic zone including the auditory areas as identified by experienced epileptologists.

Evaluation of language laterality

For each patient, hemispheric specialization of language functions was determined by a trained clinician on the basis of several clinical assessments. Those included 1) the correlation of language deficits during seizure and postictal periods, 2) a functional mapping of the regions associated to language impairment using direct electrical stimulations, and 3) a functional mapping of the regions producing gamma (>40 Hz) activity during a picture naming task [53]. Additionally, functional MRI, handedness, neuropsychological data, and, if necessary, a Wada test were also available to determine the hemispheric lateralization of language functions.

On the basis of this clinical information, we were able to classify patients into 3 groups: typical left-lateralized, atypical right-lateralized, and atypical complex (i.e., bilateral organization of language functions). For this latter group, it was difficult to determine for the different regions implicated in language processing (auditory, temporal, frontal) whether they had a typical or atypical organization.

SEEG method

A full description of the SEEG method is provided in S1 Text. Briefly, SEEG is a type of presurgical investigation based on implantation of multiple intracerebral electrodes, suitable for all types of drug-resistant epilepsies. There is no "standard" electrode implantation, the position of electrodes being chosen according to individual clinical characteristics. The implantation is performed according to the Bancaud and Talairach stereotactic method [54], with most of the electrodes implanted orthogonally through the double talairach grid.

Patients were implanted with an average of 11 (range [3–20]) depth electrodes (0.8 mm), composed of 5–15 contacts. Contacts were 2 mm long and spaced by 1.5 mm. The number of contacts per patient was on average of 138 (range [45–256]). Out of 96 patients, 45 were implanted in only one of the investigated regions. Only 3/96 patients were implanted bilaterally, including 1 in the associative cortex.

Neural recordings were performed between 4 to 9 days after the implantation procedure. No sedation or analgesics drugs were used, and antiepileptic drugs were partially or completely withdrawn. Recordings were always acquired after more than 4 hours to the last seizure.

Stimuli and paradigm

Two types of auditory stimuli were presented to patients in 2 separate sessions: 1) 30-ms–long pure tones, presented binaurally at 500 Hz or 1 kHz (with a linear rise and fall time of 0.3 ms) 110 times each, with an ISI of 1,030 (±200) ms; and 2) /ba/ or /pa/ syllables, pronounced by a French female speaker (Fig 4, top) and presented binaurally 250 times each, with an ISI of 1,030 (±200) ms. These stimuli were designed for a clinical purpose in order to functionally map the auditory cortex [30,38,39]. During the 2 recording sessions, patients laid comfortably in a chair in a sound attenuated room and listen passively to the stimuli. Auditory stimuli were delivered from loudspeakers in front of the patients at a comfortable volume. Stimuli were presented in a pseudorandom order at a 44-kHz rate using E-prime 1.1 (Psychology Software Tools Inc., Pittsburgh, PA, USA).

Functional localization of the auditory areas

Depth electrodes containing 5–15 contacts (see S1 Text) were used to perform the functional stereotactic exploration. The locations of the electrode implantations were determined solely on clinical grounds. To determine which auditory areas had been implanted, we relied on a functional localizer [30,38,39,42]. For each patient, AEPs in response to pure tones (500 Hz and 1 kHz merged) were used to functionally delineate the different auditory areas and to select the most relevant electrode contacts. AEPs were averaged over trials after epoching (−200 to 635 ms). A baseline correction was applied on each trial by computing a z-score relative to the −150 ms to 50 ms prestimulus time period. Epochs with artifacts and epileptic spikes were discarded by visual inspection prior to being averaged over trials. All contacts that elicited no significant response (<40 μV) were discarded.

The ROIs were functionally defined based on the presence of specific electrophysiological markers in the AEPs (early P20/N30, N/P50, and N/P 60–100) for PAC, SAC, and AAC (Fig 1) [30,38,39,42]. Among the 96 patients, respectively, 39, 40, and 12 had (at least) a contact in left PAC, SAC, and AAC, and 27, 29, and 12 had a contact in right PAC, SAC, and AAC. For each patient and ROI, the most responsive contact (i.e., the contact with the largest AEP) was selected for subsequent analyses when multiple contacts were present in the functional ROI. In a complementary set of analysis centered on AAC, we exploited all the contacts present in the ROI to perform an unsupervised clustering analysis.

SEEG recordings

SEEG signals were recorded at a sampling rate of 1,000 Hz using a 256-channel BrainAmp amplifier system (Brain Products GmbH, Munich, Germany) and bandpass filtered between 0.3 and 500 Hz. A scalp electrode placed in Fz was used as the recording reference. SEEG data were epoched between −5 s to 5 s relative to stimulus onset (either pure tones or syllables). Such a long temporal window for epoching allowed a more precise frequency resolution for time-frequency analysis. Epochs with artifacts and epileptic spikes were discarded by visual inspection. Data were referenced into a bipolar montage by subtracting activity recorded at each contact of interest from activity acquired at its closest neighbor site within the same electrode.

ITPC analysis

Trial-by-trial time-frequency analysis was carried out in a frequency range of 2–250 Hz (logarithmically spaced). The time-resolved spectral decomposition was performed by applying a Morlet wavelet transform to the data using the MNE-python function *time_frequency.tfr_morlet* (n_cycles = 7 and frequency steps = 100) [55]. This function also returns the ITPC, which is an estimate across trials of the time-frequency profile of the evoked activity. This measure quantifies the stereotypicality (i.e., consistency across trials) of the response per frequency, which highlights the intrinsic co-occurring time constants comprising the dynamical evoked response. For each time and frequency point, an ITPC value close to 0 reflects low phase consistency across trials, whereas an ITPC value of 1 reflects a perfect phase consistency across trials. The ITPC spectrum was then computed by averaging over time the ITPC values within a time window of interest designed to encompass one oscillatory cycle. Hence, the time window varies across frequencies (for example, 0–500 ms at 2 Hz or 0–20 ms at 50 Hz; see black overlays in Fig 2A).

To investigate the frequencies at which the evoked activity was maximal and to be able to compare them across ROIs and patients, the resulting ITPC spectra were normalized across frequencies (z-score per patient and ROI). Finally, the 2 main peaks of each ITPC spectrum (each ROI and patient) were extracted. Automatic peak detection of the 2 highest noncontiguous local maxima was performed by use of the function find_peaks from the python package scipy.signal (minimal distance between peaks = 22 points; prominence of peaks = 0.01).

LOOCV

An LOOCV was performed on all ITCP spectra from (left and right) AAC to assess whether the implantation hemisphere (left or right) could be predicted from the spectral response profile of the region to a brief auditory stimulation. In brief, for each patient implanted in AAC, we estimated the MSE (i.e., the error of fit) between the individual ITPC spectrum and both (1) the group-level ITPC pattern from all other patients implanted in the same hemisphere and (2) the group-level ITPC pattern from all patients implanted in the opposite hemisphere. For each of these 2 measures, we determined an index of similarity, computed as the squared inverse of the MSE. Finally, we compared these 2 indices of similarity to evaluate whether a patient's ITPC spectrum was more similar to the group-level ITPC pattern of the same hemisphere or to the group-level ITPC pattern of the opposite hemisphere. Identical values indicate equal similarity to both group-level ITPC patterns.

NMF

All possible bipolar montages (i.e., 245) from the electrodes (i.e., 24) implanted in AAC were preprocessed using the same analysis pipeline as previously described. In brief, each bipolar montage's time series was band-passed, epoched, transformed into ITPC, z-scored, and then time-averaged, resulting in individual ITPC spectra.

An NMF was conducted simultaneously on all resulting ITPC spectra. This clustering method was used to uncover prototypical ITPC patterns in an unsupervised manner. The non-negative input matrix V corresponding to the ITPC spectra [m frequencies × *n* contacts] is approximated as the matrix product of 2 non-negative matrix factors W [m × k] and H [k × n] (with k corresponding to the number of components of the decomposition) by optimizing the distance between V and W × H by using the squared Frobenius norm. The resulting matrix W contains a set of basis vectors that are linearly combined using the coefficients in H to represent the input data V. W gives meaningful "cluster centroids," which are prototypical ITPC patterns [56]. H represents the weight matrix, or clusters membership. It is the

estimate of the relative contribution of each component to individual ITPC spectra. The function decomposition.NMF() from the scikit learn python package was used to compute the NMF with a number of 4 components (k = 4) [57].

The proportion of variance explained (r2) by the 4 components of the NMF was computed in a cross-validation scheme. The data set was split into 2 subsets, training and testing, allowing fitting the NMF model on the training data (80% of the entire data set) in order to make predictions on the test data (20%).

Statistical procedures

All analyses were performed at the level of individual electrodes contacts (bipolar montages) before applying standard nonparametric statistical tests at the group level (unpaired nonparametric Wilcoxon–Mann–Whitney tests or parametric *t* tests).

Code availability

Codes to reproduce the results and figures of this manuscript are available on GitHub: https://github.com/DCP-INS/asymmetric-sampling.

Supporting information

S1 Fig. Individual examples of ITPC spectra in response to pure tones (0.5 kHz and 1 kHz merged) from 18 patients implanted in different auditory areas (yellow: Left hemisphere; blue: Right hemisphere). Dashed vertical lines indicate the 2 highest noncontiguous local maxima (black: first peak; gray: second peak). ITPC, intertrial phase coherence.
(TIF)

S2 Fig. ITPC spectra in response to (A) 0.5 kHz and (B) 1 kHz pure tones in the different auditory areas. Interhemispheric comparison of the ITPC spectra in PAC, SAC, and AAC. Shaded areas indicate SEM. AAC, association auditory cortex; ITPC, intertrial phase coherence; PAC, primary auditory cortex; SAC, secondary auditory cortex.
(TIF)

S1 Text. Clinical and methodological information on SEEG recordings. SEEG, stereotactic electroencephalography.
(DOCX)

Author Contributions

Conceptualization: Jérémy Giroud, Benjamin Morillon.

Data curation: Agnès Trébuchon, Patrick Marquis, Catherine Liegeois-Chauvel.

Supervision: Benjamin Morillon.

Writing – original draft: Jérémy Giroud.

Writing – review & editing: Jérémy Giroud, Agnès Trébuchon, Daniele Schön, Patrick Marquis, Catherine Liegeois-Chauvel, David Poeppel, Benjamin Morillon.

References

1. Hickok G, Poeppel D. The cortical organization of speech processing. Nat Rev Neurosci. 2007; 8: 393–402. https://doi.org/10.1038/nrn2113 PMID: 17431404

2. Zatorre RJ, Gandour JT. Neural specializations for speech and pitch: moving beyond the dichotomies. Philos Trans R Soc Lond B Biol Sci. 2008; 363: 1087–1104. https://doi.org/10.1098/rstb.2007.2161 PMID: 17890188

3. McGettigan C, Scott SK. Cortical asymmetries in speech perception: what's wrong, what's right and what's left? Trends Cogn Sci (Regul Ed). 2012; 16: 269–276. https://doi.org/10.1016/j.tics.2012.04.006 PMID: 22521208

4. Dorsaint-Pierre R, Penhune VB, Watkins KE, Neelin P, Lerch JP, Bouffard M, et al. Asymmetries of the planum temporale and Heschl's gyrus: relationship to language lateralization. Brain. 2006; 129: 1164–1176. https://doi.org/10.1093/brain/awl055 PMID: 16537567

5. Hutsler J, Galuske RAW. Hemispheric asymmetries in cerebral cortical networks. Trends Neurosci. 2003; 26: 429–435. https://doi.org/10.1016/S0166-2236(03)00198-X PMID: 12900174

6. Poeppel D. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Commun. 2003; 41: 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

7. Chait M, Greenberg S, Arai T, Simon JZ, Poeppel D. Multi-time resolution analysis of speech: evidence from psychophysics. Front Neurosci. 2015; 9: 214. https://doi.org/10.3389/fnins.2015.00214 PMID: 26136650

8. Sanders LD, Poeppel D. Local and global auditory processing: behavioral and ERP evidence. Neuropsychologia. 2007; 45: 1172–1186. https://doi.org/10.1016/j.neuropsychologia.2006.10.010 PMID: 17113115

9. Teng X, Tian X, Poeppel D. Testing multi-scale processing in the auditory system. Sci Rep. 2016; 6: 34390. https://doi.org/10.1038/srep34390 PMID: 27713546

10. Ueda K, Nakajima Y, Ellermeier W, Kattner F. Intelligibility of locally time-reversed speech: A multilingual comparison. Sci Rep. 2017; 7: 1782. https://doi.org/10.1038/s41598-017-01831-z PMID: 28496124

11. Luo H, Poeppel D. Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. Front Psychol. 2012; 3: 170. https://doi.org/10.3389/fpsyg.2012.00170 PMID: 22666214

12. Boemio A, Fromm S, Braun A, Poeppel D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci. 2005; 8: 389–395. https://doi.org/10.1038/nn1409 PMID: 15723061

13. Teng X, Tian X, Rowland J, Poeppel D. Concurrent temporal channels for auditory processing: Oscillatory neural entrainment reveals segregation of function at different scales. PLoS Biol. 2017; 15: e2000812. https://doi.org/10.1371/journal.pbio.2000812 PMID: 29095816

14. Giraud A-L, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci. 2012; 15: 511–517. https://doi.org/10.1038/nn.3063 PMID: 22426255

15. Saoud H, Josse G, Bertasi E, Truy E, Chait M, Giraud A-L. Brain-speech alignment enhances auditory cortical responses and speech perception. J Neurosci. 2012; 32: 275–281. https://doi.org/10.1523/JNEUROSCI.3970-11.2012 PMID: 22219289

16. Giraud A-L, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron. 2007; 56: 1127–1134. https://doi.org/10.1016/j.neuron.2007.09.038 PMID: 18093532

17. Morillon B, Lehongre K, Frackowiak RSJ, Ducorps A, Kleinschmidt A, Poeppel D, et al. Neurophysiological origin of human brain asymmetry for speech and language. Proc Natl Acad Sci USA. 2010; 107: 18688–18693. https://doi.org/10.1073/pnas.1007189107 PMID: 20956297

18. Morillon B, Liégeois-Chauvel C, Arnal LH, Bénar C-G, Giraud A-L. Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. Front Psychol. 2012; 3: 248. https://doi.org/10.3389/fpsyg.2012.00248 PMID: 22833730

19. Lehongre K, Morillon B, Giraud A-L, Ramus F. Impaired auditory sampling in dyslexia: further evidence from combined fMRI and EEG. Front Hum Neurosci. 2013; 7: 454. https://doi.org/10.3389/fnhum.2013.00454 PMID: 23950742

20. Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, et al. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. 2013; 11: e1001752. https://doi.org/10.1371/journal.pbio.1001752 PMID: 24391472

21. Belin P, Zilbovicius M, Crozier S, Thivard L, Fontaine A, Masure MC, et al. Lateralization of speech and auditory temporal processing. J Cogn Neurosci. 1998; 10: 536–540. https://doi.org/10.1162/089892998562834 PMID: 9712682

22. Zatorre RJ, Belin P. Spectral and temporal processing in human auditory cortex. Cereb Cortex. 2001; 11: 946–953. https://doi.org/10.1093/cercor/11.10.946 PMID: 11549617

23. Abrams DA, Nicol T, Zecker S, Kraus N. Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. J Neurosci. 2008; 28: 3958–3965. https://doi.org/10.1523/JNEUROSCI.0187-08.2008 PMID: 18400895

24. Schönwiesner M, Rübsamen R, von Cramon DY. Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. Eur J Neurosci. 2005; 22: 1521–1528. https://doi.org/10.1111/j.1460-9568.2005.04315.x PMID: 16190905

25. Overath T, Kumar S, von Kriegstein K, Griffiths TD. Encoding of spectral correlation over time in auditory cortex. J Neurosci. 2008; 28: 13268–13273. https://doi.org/10.1523/JNEUROSCI.4596-08.2008 PMID: 19052218

26. Thompson EC, Woodruff Carr K, White-Schwoch T, Tierney A, Nicol T, Kraus N. Hemispheric asymmetry of endogenous neural oscillations in young children: implications for hearing speech in noise. Sci Rep. 2016; 6: 19737. https://doi.org/10.1038/srep19737 PMID: 26804355

27. Obleser J, Eisner F, Kotz SA. Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. J Neurosci. 2008; 28: 8116–8123. https://doi.org/10.1523/JNEUROSCI.1290-08.2008 PMID: 18685036

28. Flinker A, Doyle WK, Mehta AD, Devinsky O, Poeppel D. Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. Nature Human Behavior. 2019; 3: 393–405.

29. Scott SK, McGettigan C. Do temporal processes underlie left hemisphere dominance in speech perception? Brain Lang. 2013; 127: 36–45. https://doi.org/10.1016/j.bandl.2013.07.006 PMID: 24125574

30. Liégeois-Chauvel C, de Graaf JB, Laguitton V, Chauvel P. Specialization of left auditory cortex for speech perception in man depends on temporal coding. Cereb Cortex. 1999; 9: 484–496. https://doi.org/10.1093/cercor/9.5.484 PMID: 10450893

31. Trébuchon-Da Fonseca A, Giraud K, Badier J-M, Chauvel P, Liégeois-Chauvel C. Hemispheric lateralization of voice onset time (VOT) comparison between depth and scalp EEG recordings. Neuroimage. 2005; 27: 1–14. https://doi.org/10.1016/j.neuroimage.2004.12.064 PMID: 15896982

32. Jamison HL, Watkins KE, Bishop DVM, Matthews PM. Hemispheric specialization for processing auditory nonspeech stimuli. Cereb Cortex. 2006; 16: 1266–1275. https://doi.org/10.1093/cercor/bhj068 PMID: 16280465

33. Zatorre RJ. Pitch perception of complex tones and human temporal-lobe function. J Acoust Soc Am. 1988; 84: 566–572. https://doi.org/10.1121/1.396834 PMID: 3170948

34. Deco G, Jirsa VK, McIntosh AR. Emerging concepts for the dynamical organization of resting-state activity in the brain. Nat Rev Neurosci. 2011; 12: 43–56. https://doi.org/10.1038/nrn2961 PMID: 21170073

35. Womelsdorf T, Valiante TA, Sahin NT, Miller KJ, Tiesinga P. Dynamic circuit motifs underlying rhythmic gain control, gating and integration. Nat Neurosci. 2014; 17: 1031–1039. https://doi.org/10.1038/nn.3764 PMID: 25065440

36. Harquel S, Bacle T, Beynel L, Marendaz C, Chauvin A, David O. Mapping dynamical properties of cortical microcircuits using robotized TMS and EEG: Towards functional cytoarchitectonics. Neuroimage. 2016; 135: 115–124. https://doi.org/10.1016/j.neuroimage.2016.05.009 PMID: 27153976

37. Zaehle T, Lenz D, Ohl FW, Herrmann CS. Resonance phenomena in the human auditory cortex: individual resonance frequencies of the cerebral cortex determine electrophysiological responses. Exp Brain Res. 2010; 203: 629–635. https://doi.org/10.1007/s00221-010-2265-8 PMID: 20449728

38. Liegeois-Chauvel C, Musolino A, Chauvel P. Localization of the primary auditory area in man. Brain. 1991; 114 (Pt 1A): 139–151. https://doi.org/10.1093/oxfordjournals.brain.a101854

39. Liégeois-Chauvel C, Musolino A, Badier JM, Marquis P, Chauvel P. Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. Electroencephalogr Clin Neurophysiol. 1994; 92: 204–214. https://doi.org/10.1016/0168-5597(94)90064-7 PMID: 7514990

40. Moerel M, De Martino F, Formisano E. An anatomical and functional topography of human auditory cortical areas. Front Neurosci. 2014; 8: 225. https://doi.org/10.3389/fnins.2014.00225 PMID: 25120426

41. Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage. 2001; 13: 684–701. https://doi.org/10.1006/nimg.2000.0715 PMID: 11305897

42. Liégeois-Chauvel C, Lorenzi C, Trébuchon A, Régis J, Chauvel P. Temporal envelope processing in the human left and right auditory cortices. Cereb Cortex. 2004; 14: 731–740. https://doi.org/10.1093/cercor/bhh033 PMID: 15054052

43. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, et al. A multi-modal parcellation of human cerebral cortex. Nature. 2016; 536: 171–178. https://doi.org/10.1038/nature18933 PMID: 27437579

44. Tzourio-Mazoyer N, Perrone-Bertolotti M, Jobard G, Mazoyer B, Baciu M. Multi-factorial modulation of hemispheric specialization and plasticity for language in healthy and pathological conditions: A review. Cortex. 2017; 86: 314–339. https://doi.org/10.1016/j.cortex.2016.05.013 PMID: 27321148

45. Schroeder CE, Lakatos P. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 2009; 32: 9–18. https://doi.org/10.1016/j.tins.2008.09.012 PMID: 19012975

46. Schönwiesner M, Zatorre RJ. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc Natl Acad Sci USA. 2009; 106: 14611–14616. https://doi.org/10.1073/pnas.0907682106 PMID: 19667199

47. Norman-Haignere S, Kanwisher NG, McDermott JH. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. Neuron. 2015; 88: 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035 PMID: 26687225

48. Hyafil A, Fontolan L, Kabdebon C, Gutkin B, Giraud A-L. Speech encoding by coupled cortical theta and gamma oscillations. elife. 2015; 4: e06213. https://doi.org/10.7554/eLife.06213 PMID: 26023831

49. Keitel A, Gross J, Kayser C. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biol. 2018; 16: e2004473. https://doi.org/10.1371/journal.pbio.2004473 PMID: 29529019

50. Zatorre RJ, Belin P, Penhune VB. Structure and function of auditory cortex: music and speech. Trends Cogn Sci (Regul Ed). 2002; 6: 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7

51. Sammler D, Grosbras M-H, Anwander A, Bestelmeyer PEG, Belin P. Dorsal and ventral pathways for prosody. Curr Biol. 2015; 25: 3079–3085. https://doi.org/10.1016/j.cub.2015.10.009 PMID: 26549262

52. Ghitza O. Acoustic-driven delta rhythms as prosodic markers. Lang Cogn Neurosci. 2017; 32: 545–561. https://doi.org/10.1080/23273798.2016.1232419

53. Alario FX, Ferrand L. A set of 400 pictures standardized for French: norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. Behav Res Methods Instrum Comput. 1999; 31: 531–552. https://doi.org/10.3758/bf03200732 PMID: 10502875

54. Talairach J, Bancaud J, Bonis A, Szikla G, Tournoux P. Functional stereotaxic exploration of epilepsy. Stereotact Funct Neurosurg. 1962; 22: 328–331. https://doi.org/10.1159/000104378 PMID: 13984743

55. Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MNE software for processing MEG and EEG data. Neuroimage. 2014; 86: 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027 PMID: 24161808

56. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401: 788–791. https://doi.org/10.1038/44565 PMID: 10548103

57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011.

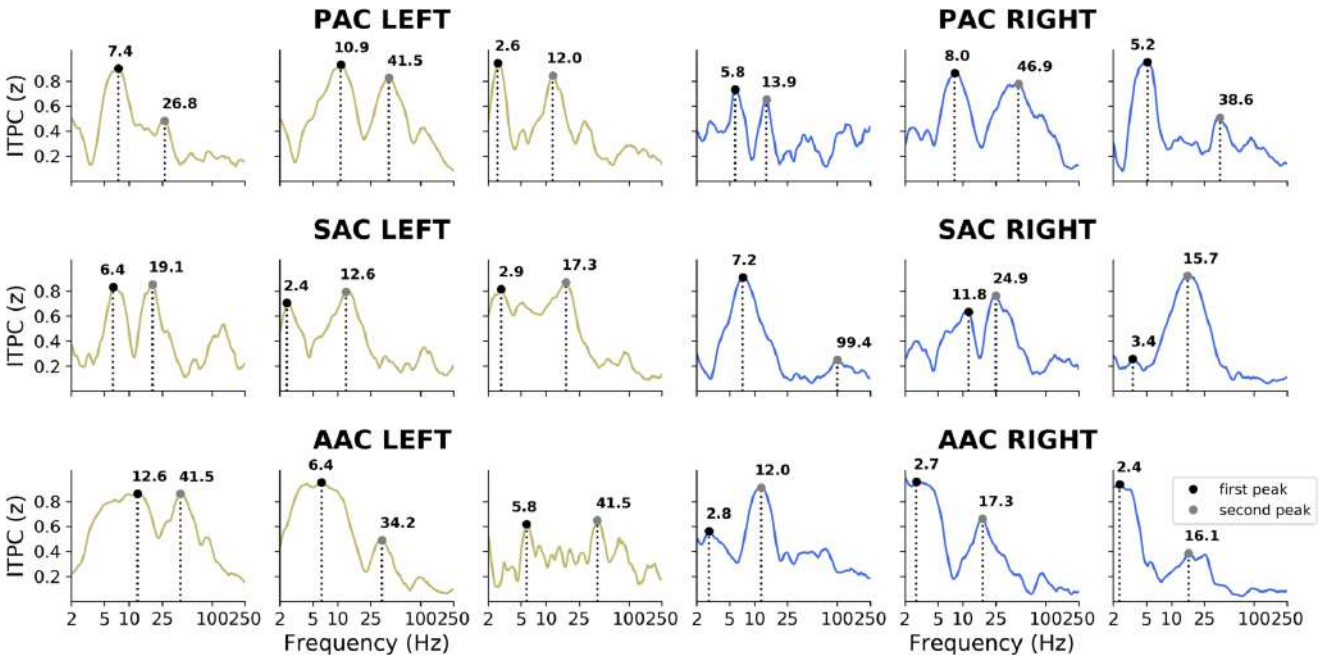


Fig S1. ***Individual examples of ITPC spectra in response to pure tones*** (0.5 kHz and 1 kHz merged) from 18 patients implanted in different auditory areas (yellow: Left hemisphere; blue: Right hemisphere). Dashed vertical lines indicate the 2 highest noncontiguous local maxima (black: first peak; gray: second peak). ITPC, intertrial phase coherence.

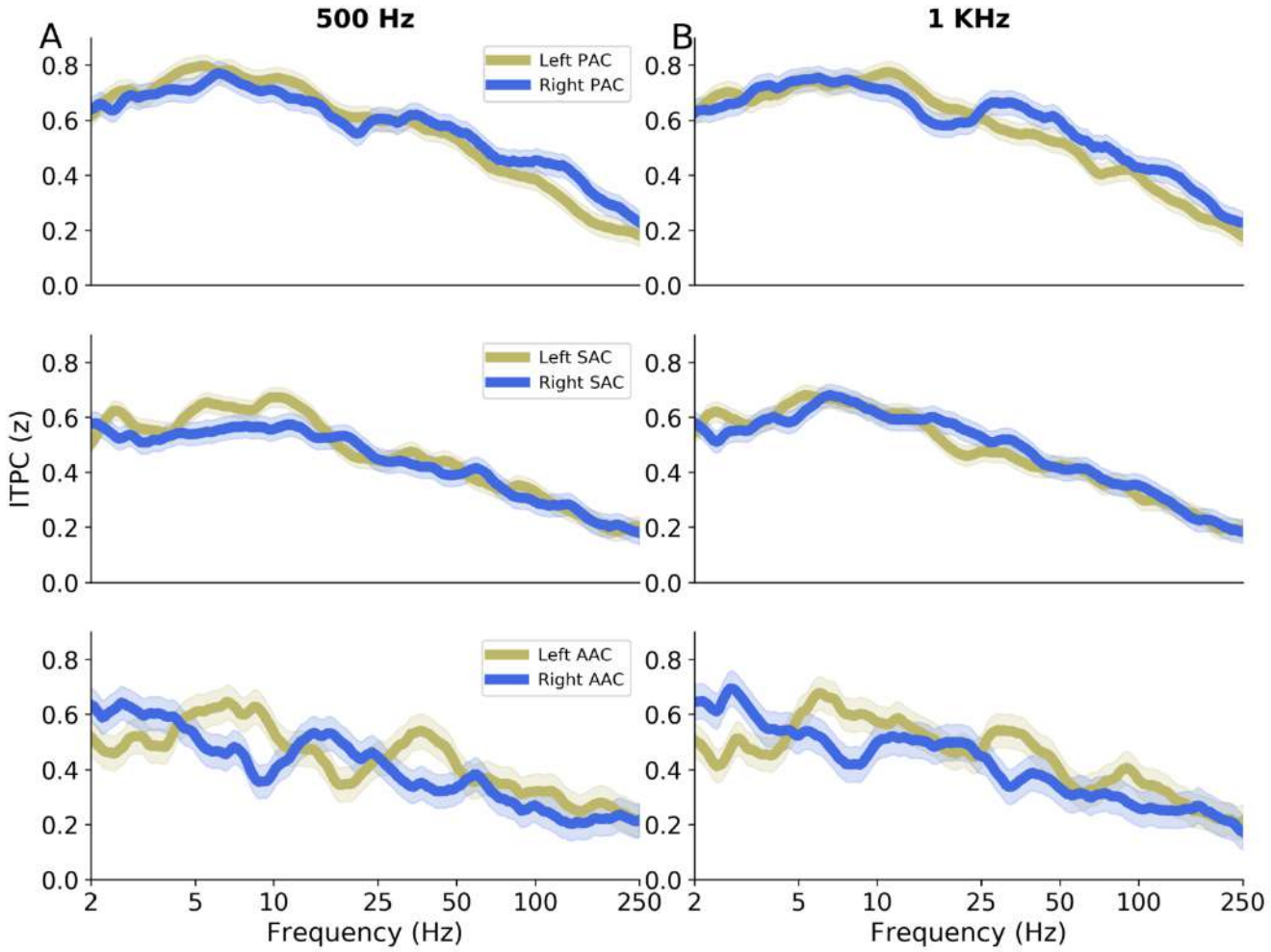


Fig S2. ITPC spectra in response to (A) 0.5 kHz and (B) 1 kHz pure tones in the different auditory areas. Interhemispheric comparison of the ITPC spectra in PAC, SAC, and AAC. Shaded areas indicate SEM. AAC, association auditory cortex; ITPC, intertrial phase coherence; PAC, primary auditory cortex; SAC, secondary auditory cortex.

Supporting information

The stereotactic EEG (SEEG) method. SEEG has been for fifty years (developed by Jean Talairach) a type of presurgical investigation suitable for all types of drug-resistant epilepsies. SEEG is based on implantation of multiple intracerebral electrodes in order to record and map the epileptogenic and propagation networks. The SEEG method includes recording of spontaneous seizures, recording of neural activity during sensory stimulation, and direct cortical electrical stimulation to trigger seizures (part of the localization process) and perform functional mapping in the same areas (to evaluate possible contraindications due to sensory or cognitive impairment such as language deficits).

Strategy of implantation. There is no "standard" electrode implantation. Implantation strategy depends only on the hypotheses made about the anatomical location of the epileptogenic zone (EZ), with the aim of defining subsequent cortectomy. The hypotheses of the likely EZ localization are based on (phase I) non-invasive pre-surgical assessment including detailed clinical history, surface video-electroencephalographic (EEG) recording, MRI, and 18FDG-PET scanner. Electrode positions are therefore not standardized across patients, but chosen according to individual clinical characteristics.

One of the most frequent regions to be explored is the perisylvian region in order to know for instance if the patient's epilepsy is temporal, temporo-perisylvian, or perisylvian (the main cause of surgical failure in temporal epilepsy is a misdiagnosis of perisylvian epilepsies). Perisylvian epilepsy can be localized in insular cortex, frontal opercular, parietal opercular, temporal opercular cortices and superior temporal and supramarginal gyri. In this region, electrodes are implanted orthogonally to the cortical surface in order to record along one single electrode both posterior insula, tip of Heschl's gyrus and planum temporale, for example. Another trajectory is to implant more anteriorly the superior temporal gyrus and the ventral insula. The perisylvian region is also implanted in the hypothesis of inferior parietal lobule, pericentral, or ventral prefrontal/premotor epilepsy. The transverse gyrus (Heschl's gyrus, which includes the auditory cortex) is critically important because of its connections with lower central regions and the inferior frontal gyrus: it is a pathway through which seizures from the temporal pole and the anterior superior temporal gyrus generalize.

Implantation procedure. The implantation of electrodes is performed according to the Bancaud and Talairach stereotactic method [1]. The stereotaxic localization and the implantation of the intracerebral electrodes are done at the same time. The Leksell frame is fixed percutaneously, and MRI and cerebral angiography are obtained. Then, a catheter cerebral angiogram is obtained in telemetric and stereotaxic conditions, and trajectories are planned by looking at both 3D T1-weighted (T1W) MR and brain vasculature. All the anatomic information is merged in the stereotaxic space. Most of the electrodes are orthogonally implanted through the double talairach grid. The entry point of the electrodes is determined on the basis of the disposition of the vessels in the area. Finally, electrodes are implanted percutaneously by twist drill, and fixed using anchor bolts.

Post-implantation procedure. A post-operative electrode implantation computerized tomography (CT) scan is done to verify the absence of bleeding and the location of each recording lead. Following this, intracerebral electrodes are removed and an MRI scan is performed, permitting to visualise each electrode's trajectory. Finally, CT-scan/MRI data fusion is performed to locate each contact along the electrode trajectory. The fusion can be done with several software: BrainVISA/Anatomist, medInria, Gardel [2–4].

Potential impact of drugs on neural recordings. Neural recordings are performed between 4 to 9 days after the implantation procedure. No sedation or analgesics drugs are used. In general, a partial or complete withdrawal of antiepileptic drugs is done prior to the beginning of the exploration. However, the level of medication is adjusted on an individual basis, as a function of the seizure type. Recordings are always acquired after more than 4 hours to the last seizure.

Localisation of electrode contacts in the current dataset. The 96 patients of our study were implanted with an average of 11 (median: 11; range [3-20]) depth electrodes (0.8 mm; Alcis, Besançon, France). Electrodes had between 5 and 15 contacts. Contacts were 2 mm long and spaced from each other by 1.5 mm. The number of contacts per patient was on average 138 (median: 124; range [45-256]). The number of contacts per region depended on the implantation location and the nature of the region. More specifically, more associative regions are larger than primary ones. Thus, the primary auditory cortex was implanted with up to 3 contacts, the secondary cortex with up to 5 contacts, and the associative cortex with up to 15 contacts. Out of the 96 implanted patients, 45 were implanted in only one of the ***investigated regions.*** Most of the patients implanted in more than one region were implanted in the primary and secondary auditory cortex of the same hemisphere. Only 3 patients were implanted bilaterally, and only 1 patient was implanted bilaterally in the associative cortex.

REFERENCES

1. Talairach J, Bancaud J, Bonis A, Szikla G, Tournoux P. Functional stereotaxic exploration of epilepsy. Stereotact Funct Neurosurg. 1962;22: 328–331. doi:10.1159/000104378

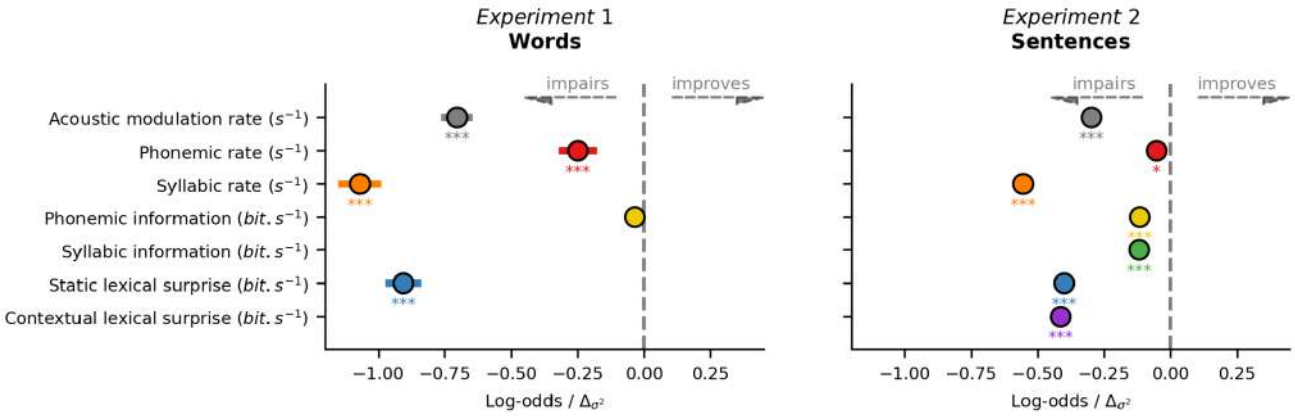
2. Rivière D, Geffroy D, Denghien I, Souedet N, Cointepas Y. BrainVISA: an extensible software environment for sharing multimodal neuroimaging data and processing tools. Neuroimage. 2009;47: S163. doi:10.1016/S1053-8119(09)71720-3

3. medInria. In: med.inria [Internet]. [cited 6 Feb 2019]. Available: http://med.inria.fr/

4. Medina Villalon S, Paz R, Roehri N, Lagarde S, Pizzo F, Colombet B, et al. EpiTools, A software suite for presurgical brain mapping in epilepsy: Intracerebral EEG. J Neurosci Methods. 2018;303: 7–15. doi:10.1016/j.jneumeth.2018.03.018

Study 2

The channel capacity of multilevel linguistic features constrains speech comprehension



The channel capacity of multilevel linguistic features constrains speech comprehension

Jérémy Giroud[1*], Jacques Pesnot Lerousseau[1], François Pellegrino[2], Benjamin Morillon[1,3]

[1] Aix Marseille Univ, Inserm, INS, Inst Neurosci Syst, Marseille, France

[2] Laboratoire Dynamique du Langage UMR 5596, CNRS, University of Lyon, 14 Avenue Berthelot, 69007 Lyon, France.

[3] Senior authorship

* corresponding author : jeremy.giroud@univ-amu.fr

Keywords: accelerated speech, syllabic rate, information rate, phonemic rate, behavior, humans, auditory psychophysics, gating paradigm

Corresponding Author and Lead Contact: Jérémy Giroud, Aix-Marseille Univ, INS, Inst Neurosci Syst, Marseille, France; jeremy.giroud@univ-amu.fr

Conflict of interests: The authors declare no competing interests.

Acknowledgments: We thank all participants; Johanna Nicolle, François-Xavier Alario and all the colleagues from the DCP team at the Institut de Neurosciences des Systèmes for useful discussions; Yannick Jadoul for help with the Parselmouth python package and Ting Qian from FindingFive for extensive assistance and advice.

Funding sources: ANR-20-CE28-0007-01 (to B.M), ANR-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), ANR-17-EURE-0029 (NeuroMarseille), the French government under the Programme «Investissements d'Avenir», the Initiative d'Excellence d'Aix-Marseille Université (A*MIDEX, AMX-19-IET-004), la Ligue Française Contre l'Épilepsie (LFCE, to J.G).

Author contributions: Conceptualization J.G., F.P., B.M; Data curation J.G; Formal Analysis J.G., J.P.L.; Funding acquisition B.M; Investigation J.G., J.P.L., F.P., B.M; Methodology J.G., J.P.L., F.P., B.M; Project administration B.M.; Resources B.M.; Supervision B.M.; Software J.G., J.P.L.; Validation F.P., B.M.; Visualization J.G., J.P.L. and B.M.; Writing – original draft J.G.; Writing – review & editing J.G, J.P.L., F.P., B.M..

Abstract

Humans are expert at processing speech but how this feat is accomplished remains a major question in cognitive neuroscience. Capitalizing on the concept of channel capacity, we developed a unified measurement framework to investigate the respective influence of seven linguistic features on speech comprehension, encompassing acoustic, sub-lexical, lexical and supra-lexical levels of description. We show that comprehension is independently impacted by all these features, but at varying degrees and with a clear dominance of the syllabic rate. Comparing comprehension of French words and sentences further reveals that when supra-lexical contextual information is present, the impact of all other features is dramatically reduced. Finally, we estimated the channel capacity associated with each linguistic feature and compared them with their generic distribution in natural speech. Our data point towards supra-lexical contextual information as the feature limiting the flow of natural speech. Overall, this study reveals how multilevel linguistic features constrain speech comprehension.

Introduction

Humans are remarkably successful at quickly and effortlessly extracting meaning from spoken language. The classical method to study this ability and identify its processing steps is to reveal the constraints that limit speech comprehension. For example, the fact that speech comprehension drops when more than ~12 syllables per second are presented has been interpreted as evidence that at least one processing step concerns syllables extraction (Ghitza, 2013; Giraud & Poeppel, 2012; Versfeld & Dreschler, 2002). As language processing involves distinct representational and temporal scales, it is usually decomposed into co-existing levels of information, estimated with distinct linguistic features, from acoustic to supra-lexical (Christiansen & Chater, 2016; Gregory Hickok & Poeppel, 2007; Rosen, 1992). However, most studies only investigated a single linguistic feature and, as a consequence, a complete picture of which processes underlie speech comprehension is still lacking. This is because there exists no common theoretical framework and no unique experimental paradigm to compare multiple linguistic features at the same time. Among the existing experimental paradigms, accelerating speech to generate adverse and challenging comprehension situations is a common approach. However, when speech is artificially time-compressed (Dupoux & Green, 1997; Foulke & Sticht, 1969; Garvey, 1953), all linguistic features are impacted by the modification, making it impossible to disentangle their unique impact on behavioral performance. It thus remains unknown whether the syllabic rate actually constrains comprehension, whether it is the phonemic rate or any other rate, or whether bottlenecks are present at different levels of processing.

To solve this problem, we propose to rely on a concept inherited from information theory (C. E. Shannon, 1948), channel capacity, and to carefully orthogonalize multiple linguistic features to reveal their unique contribution to speech comprehension. The processing of each linguistic feature can be modeled as a transfer of information through a dedicated channel. Channel capacity is defined as the maximum rate at which information can be transmitted. Thanks to this approach, we identified and compared in a unique paradigm the potential impact of acoustic, sub-lexical, lexical and supra-lexical linguistic features on speech comprehension.

First, speech is an acoustic signal characterized by a prominent peak in its envelope modulation spectrum, around 4-5 Hz, a feature shared across languages (Ding et al., 2017; Varnet et al., 2017). This *acoustic modulation rate* approximates the *syllabic rate* of the speech stream (David Poeppel & Assaneo, 2020), which happens at around 2.5 – 8 syllables per second in natural settings (Coupé et al., 2019; Kendall, 2013; Pellegrino et al., 2011). The acoustic modulation rate can serve as an acoustic guide for parsing syllables (Mermelstein, 1975). In addition to these, comprehension depends on the linguistic coding of phonemic details, necessitating parsing speech at the *phonemic rate* (Ghitza, 2011; Giraud & Poeppel, 2012; Hyafil et al., 2015; Peelle & Davis, 2012; David Poeppel, 2003; Stevens, 2002). We thus estimated three speech rates, the raw *acoustic modulation rate*, and the linguistically-motivated *syllabic and phonemic rates*.

Second, syllabic and phonemic sub-lexical units carry linguistic information. A description of speech in terms of linguistic information rates rather than speech rates could be more appropriate to understand how language is processed (Coupé et al., 2019; Pellegrino et al., 2011; Reed & Durlach, 1998). Moreover, the information rate (in bits/s), rather than an absolute informational value (in bits), is a more relevant dimensional space (Coupé et al., 2019), in accordance with the fact that neurocognitive ressources are best characterized as temporal bottlenecks (Hasson et al., 2008; Honey et al., 2012; Lerner, Honey, Katkov, & Hasson, 2014; Lerner et al., 2011; Vagharchakian et al., 2012). Hence, we estimated *syllabic and phonemic informational rates*.

Third, at the lexical and supra-lexical levels, probabilistic constraints regulate language processing. It has been suggested that speech processing depends on predictive computations to guide the interpretation of incoming information. Predictions of upcoming individual words depend on both prior knowledge and contextual information (Brodbeck et al., 2018; Donhauser & Baillet, 2020; Gagnepain, Henson, & Davis, 2012; Gwilliams, Linzen, Poeppel, & Marantz, 2018; Kutas, DeLong, & Smith, 2011; Pickering & Garrod, 2007; Sohoglu, Peelle, Carlyon, & Davis, 2012). Lexical (or word) frequency, the probabilistic knowledge about word occurrences, has a strong impact on lexical access time (Brysbaert, Lange, & Wijnendaele, 2000; Ferreira, Henderson, Anes, Weeks, & McFarlane, 1996). Hence, we estimated the context-independant or *static lexical surprise rate*, i.e., the amount of unexpectedness of word occurrences per second (see Methods). Additionally, recent models based on deep neural networks exploit contextual lexical information to predict brain activity during natural speech processing (Caucheteux et al., 2021; Goldstein et al., 2020; Heilbron et al., 2020; Schrimpf et al., 2020). We used CamemBERT (L. Martin et al., 2020), a transformer model trained for the French language, to estimate the *contextual lexical surprise rate*, i.e., the lexical surprise rate predicted by the context provided by each sentence.

To reveal the efficiency of the speech comprehension system and estimate its capacity and limitations with unprecedented levels of granularity, we developed and combined three innovative experimental approaches: 1) First, we developed the *compressed speech gating paradigm*, a behavioral approach allowing an efficient estimation of the relation between time-compression and comprehension performance. For each stimulus a comprehension point could be determined, corresponding to the compression rate at which comprehension emerges. 2) Second, speech is in essence a temporal signal, and previous work has shown the relevance of considering linguistic features as a number of units communicated per unit of time (i.e., in rate, or bit/s; (Coupé et al., 2019; Pellegrino et al., 2011; Reed & Durlach, 1998). Each linguistic feature was thus expressed in a number of units per second. With such an approach, and utilizing the comprehension point as the maximum rate at which information is transmitted, the channel capacity associated with each linguistic feature can be estimated. Moreover, features can also be compared directly between one another and ranked according to the magnitude of their respective influence. 3) Third, to simultaneously estimate the impact of multiple linguistic features on comprehension capacities, we developed an original stimulus selection and orthogonalization procedure. We generated two speech corpora derived from large databases of natural stimuli and characterized them at seven linguistic levels, ranging from acoustic to supra-lexical. Thanks to a careful selection, all these features were orthogonalized across stimuli, enabling a fine-grained characterization of their respective influence on speech comprehension. The combination of these three methodological advances provides optimal conditions to investigate the linguistic features governing speech processing ability and limits.

Results from two behavioral experiments converge to show that multilevel linguistic features independently constrain speech comprehension, with the syllabic rate having the strongest impact. When supra-lexical contextual information is provided to participants, the impact of all other features is dramatically reduced. Estimating the channel capacity associated with each feature, we show in particular that comprehension drops when phonemic or syllabic rates are respectively above ~40 Hz or ~15 Hz. Finally, comparing these estimated channel capacities with the generic distribution of the linguistic features in natural speech, we find that at original speed contextual lexical information is already close to its channel capacity, which suggests that it is the main cognitive feature limiting the flow of natural speech.

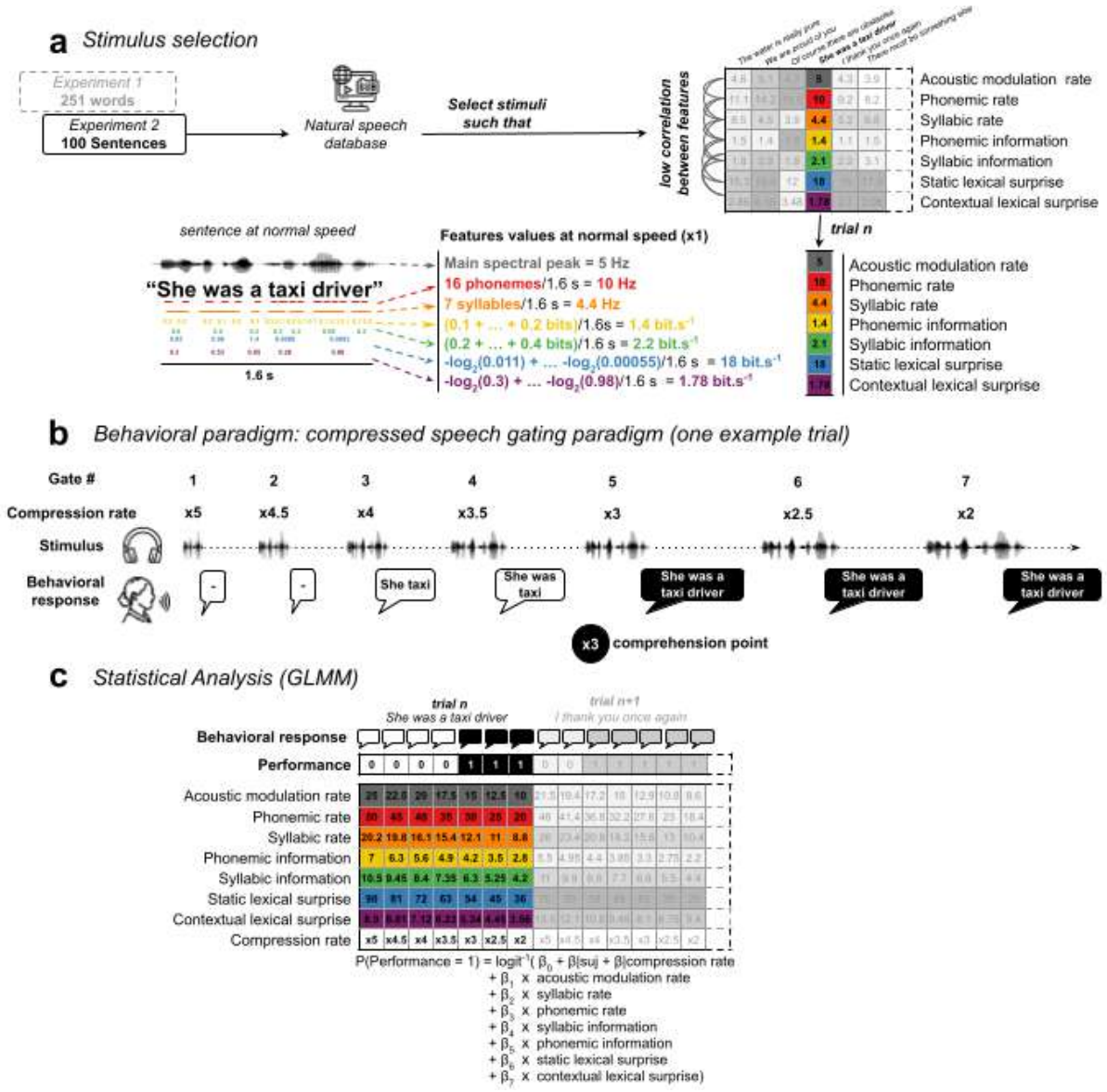


Figure 1. Experimental design and analysis pipeline. a) *Stimulus selection procedure. 251 words and 100 sentences were used in experiments 1 and 2, respectively. Word stimuli were retrieved from the French Lexique database and sentence stimuli from the Web Inventory of Transcribed and Translated Talks database. Seven linguistic features were computed for each stimulus, illustrated here for an example sentence (sentences in experiment 2 were 7-words long). Features corresponded to the acoustic modulation rate (in Hz), syllabic rate (in Hz), phonemic rate (in Hz), syllabic information rate (in bit/s), phonemic information rate (in bit/s), static lexical surprise (in bit/s) and contextual lexical surprise (in bit/s). The selection procedure ensured that low correlations (all r < 0.15) across stimuli were present between features in the selected stimulus sets (see Methods).* ***b)*** *Behavioral paradigm. A modified gating paradigm was used for both experiments. In each trial, participants were presented with time-compressed versions of the original audio stimulus, from the most to the least compressed version, and were asked to report what they heard after each audio presentation. Behavioral responses were classified into incorrect and correct responses (incorrect: white bubbles; correct: black bubbles). At each trial, a "comprehension point" (black circle) was determined. It corresponds to the compression rate at which comprehension emerged, estimated across gates with a logistic regression model (see Methods).* ***c)*** *Behavioral responses were entered into a generalized linear mixed models (GLMM) to assess the respective contribution of each feature on comprehension performance. The equation includes participants and compression rates as random effects and linguistic features as fixed effects. Entering compression rates as random effects ensured that correlations between stimuli across compression rates were controlled for in the model.*

Results

Compressed speech gating paradigm.

We collected behavioral data from two independent experiments in which 42 participants were required to understand successive time-compressed versions of either spoken monosyllabic words or sentences, respectively in Experiment 1 and 2 (Fig. 1). At each trial, the same spoken utterance was presented at decreasing compression rates ranging from unintelligible, to challenging, to intelligible. Using regression analyses, we modeled the individual comprehension performance fluctuation at the single trial level, as a function of a mixture of features encompassing the entire linguistic hierarchy from acoustic to supra-lexical levels of description. Linguistic features were chosen based on a large body of literature identifying them as influential constraints on speech comprehension (see Introduction). Our corpus selection procedure guaranteed that feature distributions selected in the final experimental material were representative of generic stimuli statistics as derived from large databases (Fig. Supp. 1a and 1d). In experiment 1, the limitations in terms of existing monosyllabic words prevented us from reaching a stimulus set in which the syllabic information rate was representative of the original database. Specifically, both the mean and variance of the distribution across stimuli differed between the original and selected stimulus sets (Fig. Sup. 1b and 1e). We thus excluded this feature from the data analyses of experiment 1. We also ensured that within the subset of selected stimuli, correlations between linguistic features were low (all $r < 0.12$; Fig. Supp. 1c and 1f), thanks to an orthogonalization procedure. This is a crucial condition to be able to determine their respective impact on speech comprehension performance. Finally, by investigating each feature in a similar measurement framework we were able to directly compare their respective impact on speech comprehension.

Compressed speech impairs speech comprehension.

Across the different compression rates, comprehension shifted from not understood (mean performance accuracy of 0.03 % and 0.1 % for experiments 1 and 2, respectively) to perfectly understood (96.3 % and 99 %), with a characteristic sigmoid function, indicating that the range of compression rates selected was well suited to investigate speech comprehension at its limits (Fig. 2). A mean performance accuracy of 50 % was observed for a compression rate of 3.5 in both experiments. At a compression rate of 5 or above, comprehension was essentially residual (< 10 %).

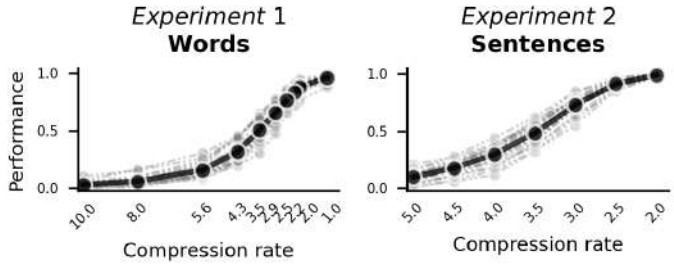


Figure 2. Comprehension performance as a function of compression rate. *Performance is expressed in proportion of correct responses. Thin dashed grey lines depict individual performance. Thick black lines indicate average performance. In experiment 1, participants were presented with the same audio stimuli (words) at ten different compression rates. In experiment 2, participants were presented with the same audio stimuli (sentences) at seven different compression rates.*

Multifactorial linguistic constraints concurrently limit speech comprehension.

We used generalized linear mixed-effects models (GLMMs) to evaluate the extent to which multiple linguistic features were predictive of behavioral performance (word or sentence comprehension). The GLMM approach enables a fine-grained characterization of the independent contributions of the different features (see Methods).

In experiment 1, a GLMM with a logit link function was conducted to model spoken word comprehension. The model included participants and compression rates as random effects and five linguistic features, acoustic modulation rate, the phonemic and syllabic rates, phonemic information rate and static lexical surprise, as fixed effects (Fig. 3, left panel; table 1; see Methods). The stimuli consisting of isolated words, no contextual lexical surprise was defined. The full model accounted for 74 % of the variance of the data. The model revealed a significant effect of the acoustic modulation rate ($\beta = -0.7 \pm 0.06$, $p < 0.001$), the phonemic rate ($\beta = -0.25 \pm 0.07$, $p = 0.001$) and the syllabic rate ($\beta = -1.07 \pm 0.08$, $p < 0.001$), indicating that they independently and additively impact comprehension. The model's coefficients read as follows: $\beta = -1.07$ means that the odds of giving a correct response are multiplied by $\exp(-1.07) \approx$ are divided by 3 for an increase of one standard deviation in syllabic rate, demonstrating the adverse impact of syllabic rate on speech comprehension. Phonemic information rate did not significantly contribute to the model ($\beta = -0.03 \pm 0.03$, $p = 0.258$). Finally, the static lexical surprise was significantly associated with listeners' speech comprehension ($\beta = -0.91 \pm 0.07$, $p < 0.001$), indicating that words' unexpectedness worsens participants' comprehension.

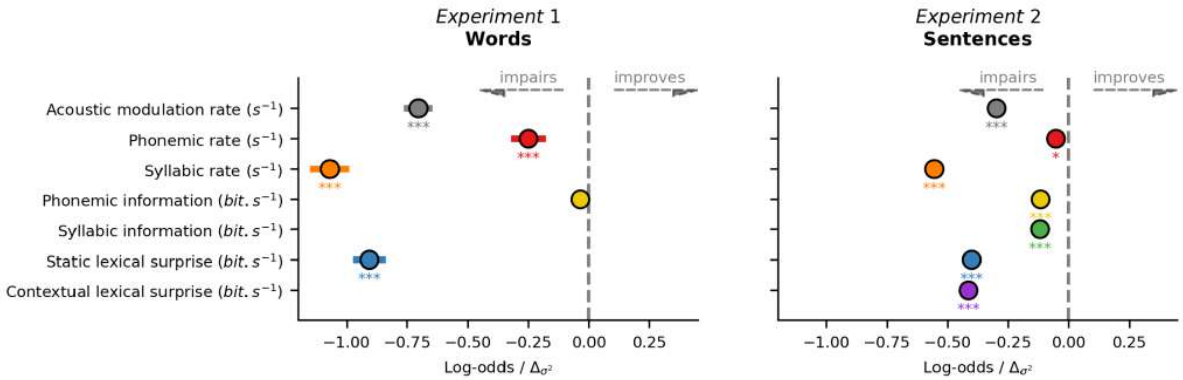


Figure 3. GLMM results. Log-odds ratios of the linguistic features included in the GLMM models in experiments 1 and 2. *Coefficients were standardized and read as follows: in experiment 1, the odds of giving a correct response are multiplied by exp(-1.12) ≈ 0.33 ≈ are divided by 3 for an increase of one standard deviation in syllabic rate (orange dot in experiment 1). In other words, an increase of one standard deviation in syllabic rate divides the odds of understanding the word by 1/exp(-1.12) ≈ 3. Negative log-odds ratios indicate a negative effect on performance. In both models, linguistic features were entered as fixed effects. Participants and compression rates were entered as random effects. *p < 0.05; ***p < 0.001. Error bars indicate standard error of the mean across participants.*

	Experiment 1 (words)					Experiment 2 (sentences)				
Fixed effects										
	Log-odds	SE	CI (95%)		p	Log-odds	SE	CI (95%)		p
Intercept	-0.02	0.32	-0.64	0.60	0.956	0.43	0.31	-0.17	1.04	0.161
Acoustic modulation rate	-0.70	0.06	-0.82	0.59	**<0.001**	-0.30	0.03	-0.35	-0.24	**<0.001**
Phonemic rate	-0.25	0.07	-0.39	-0.11	**0.001**	-0.05	0.02	-0.10	-0.01	**0.022**
Syllabic rate	-1.07	0.08	-1.23	-0.91	**<0.001**	-0.56	0.02	-0.60	-0.51	**<0.001**
Phonemic information rate	-0.03	0.03	-0.10	0.03	0.258	-0.12	0.02	-0.16	-0.07	**<0.001**
Syllabic information rate						-0.12	0.02	-0.16	-0.08	**<0.001**
Static lexical surprise	-0.91	0.07	-1.04	-0.77	**<0.001**	-0.40	0.04	-0.48	-0.32	**<0.001**
Contextual lexical surprise						-0.41	0.01	-0.44	-0.39	**<0.001**
Random effects										
σ^2	3.29					3.29				
τ_{00}	0.14 participant 0.93 compression rate					0.16 participant 0.62 compression rate				
ICC	0.25					0.19				
Number of observations										
N	10 compression rate 21 participant					7 compression rate 21 participant				
Observations	52710					14700				
Marginal R^2 / Conditional R^2	0.659 / 0.743					0.427 / 0.536				

Table 1. *General linear mixed model (GLMM) results for experiments 1 and 2 with the comprehension performance as the dependent variable.*

Holm-corrected post-hoc comparisons were performed to identify differences among selected features in modulating spoken word comprehension. Features were ordered from the most to the least influential, and compared between neighbours. The analysis revealed no significant difference between the two most influential features, syllabic rate and static lexical surprise ($\beta = -0.16$, $z = -1.58$, $p = 0.12$). In contrast, all other pairwise comparisons were significantly different (all $p < 0.05$).

In experiment 2, a GLMM with a logit link function was also used to model spoken sentences comprehension. The model included seven linguistic features as fixed effects (Fig. 3, right panel; table 1; see Methods). All linguistic features significantly contributed to the model and together explain 54 % of the variance of the data (Fig. 3, right panel; table 1). Similar to experiment 1, post-hoc comparisons were conducted to assess differences between the relative influence of each linguistic feature on sentence comprehension. The analysis showed that the syllabic rate has the largest impact on performance, with significantly more influence than contextual lexical surprise ($\beta = -0.14$, $z = -5.22$, $p < 0.001$). Conversely, the contrast between contextual and static lexical surprise rate did not reach

significance (β = -0.01, z = -0.34, p > 0.05). Whereas modulatory effect of the static lexical surprise and the acoustic modulation rate on comprehension was not significantly different (β = -0.10, z = -2.07, p > 0.05), this latter alter significantly more speech comprehension than syllabic information rate (β = -0.18, z = -4.87, p < 0.001). Finally, modulation of performance induced by syllabic information rate, phonemic information rate and phonemic rate do not significantly differ (all p > 0.41).

Adding contextual information reduces the influence of the other linguistic features.

Comparing experiments 1 and 2, we first observed a similar profile of response weights, with a larger impact of syllabic rate and static lexical surprise, a medium influence of the acoustic modulation rate, and lower weights for the other linguistic features (Fig. 3). Moreover, the addition of contextual information in experiment 2, with the presence of seven linguistically related words, whose effect was modelled in the contextual lexical surprise, led to a reduction of the influence of all other features on comprehension (sum of weights of the five linguistic features present in both experiments: -2.96 in experiment 1; -1.43 in experiment 2). Random factors had similar weights across experiments, with an expected strong influence of the compression rate on comprehension performance, and a minor effect of the inter-individual variability, indicating that our results are robust across participants.

	Experiment 1 (words)					**Experiment 2 (sentences)**				
Fixed effects										
	Log-odds	SE	CI (95%)		p	Log-odds	SE	CI (95%)		p
Intercept	3.92	0.10	3.73	4.10	**<0.001**	3.23	0.03	3.16	3.29	**<0.001**
Acoustic modulation rate	-0.19	0.03	-0.25	-0.13	**<0.001**	-0.03	0.01	-0.06	-0.01	**0.002**
Phonemic rate	-0.06	0.03	-0.12	-0.01	**0.027**	-0.03	0.01	-0.05	-0.00	**0.019**
Syllabic rate	-0.21	0.03	-0.26	-0.15	**<0.001**	-0.05	0.01	-0.08	-0.03	**<0.001**
Phonemic information rate	-0.01	0.03	-0.06	0.05	0.765	0.00	0.01	-0.02	0.02	0.989
Syllabic information rate						-0.03	0.01	-0.06	-0.01	**0.002**
Static lexical surprise	-0.20	0.03	-0.26	-0.15	**<0.001**	-0.04	0.01	-0.07	-0.02	**<0.001**
Contextual lexical surprise						-0.20	0.01	-0.22	-0.18	**<0.001**
Random effects										
σ^2	4.16					0.25				
τ_{00}	0.18 participant					0.02 participant				
ICC	0.04					0.08				
Number of observations										
N	21 participant					21 participant				
Observations	5250					2100				

Marginal R^2 / Conditional R^2	0.025 / 0.065	0.172 / 0.241

Table 2. *Linear mixed model (LMM) results for experiments 1 and 2.*

Multilevel linguistic features consistently shift the comprehension point.

Following the main GLMM analysis, we aimed at characterizing the relationship between the value of each linguistic feature at original speed (x1) – which reflects the intrinsic linguistic properties of the stimulus sets – and the comprehension point (i.e the compression rate at which participants' comprehension reaches 75 % of accuracy, see Methods). This analysis ought to confirm the individual propensity of each linguistic feature to modulate the comprehension point (see Methods). In experiment 1, a linear mixed model analysis fully reproduced the results from the main GLMM analysis (Table 2), revealing a significant impact of all features but the phonemic information rate, on comprehension (all $p < 0.05$). In experiment 2, the linear mixed model revealed that, apart from phonemic information rate, all other features significantly delayed the comprehension point (all $p < 0.05$), also confirming the previous analysis. The putative effect size associated with phonemic information rate is probably negligible, even if significance has been limited by the number of observations taken into account in this alternative model (2100 vs. 14700 behavioral responses, see Methods). Overall, these new analyses confirm the robustness of the results previously obtained with the GLMM and directly show that the linguistic properties of the non-compressed stimuli predict the maximal compression rate at which comprehension can be maintained.

The syllabic rate is the strongest determinant of speech comprehension.

To more directly visualise the data from both experiments, a complementary approach was adopted. For each compression rate, performance was first binned as a function of the syllabic rate (see Methods), as this feature had the strongest impact on performance in the two experiments (Fig. Supp. 2 and Fig. Supp. 3a). This visualisation highlights the major influence of the syllabic rate on behavioral outcome independently of the compression rate, in both experiments. Second, data were also binned as a function of the other features, after having been stratified as a function of the syllabic rate (Fig. Supp. 2 and Fig. Supp. 3). This highlights their additional impact over the major influence of syllabic rate. This visualisation enables a better grasping of the relative influence of each linguistic feature on comprehension and confirmed graphically the genuine results obtained with the more fine-grained GLMM and LMM approaches.

Estimation of the channel capacity associated with each linguistic feature.

Thanks to the compressed speech gating paradigm, we were able to derive for each feature the distribution of its values (in rate) at the comprehension point, which provided an estimation of its channel capacity (see Methods). This estimation corresponds to the value (in rate, or bit/s) at which comprehension consistently emerges. This threshold thus reflects a successful transmission of linguistic information but also determines the highest rate of information flow. As such, stimuli containing linguistic feature’s values above this threshold will exceed channel capacity leading to a drop in comprehension performance. Overall, we found that channel capacities associated with each linguistic feature investigated were on the same order of magnitude in both experiments (Fig. 4). Specifically, the estimated maximum acoustic modulation and syllabic rates were both centred around 10-15 Hz, while the phonemic rate’s channel capacity was centred around 35 Hz.

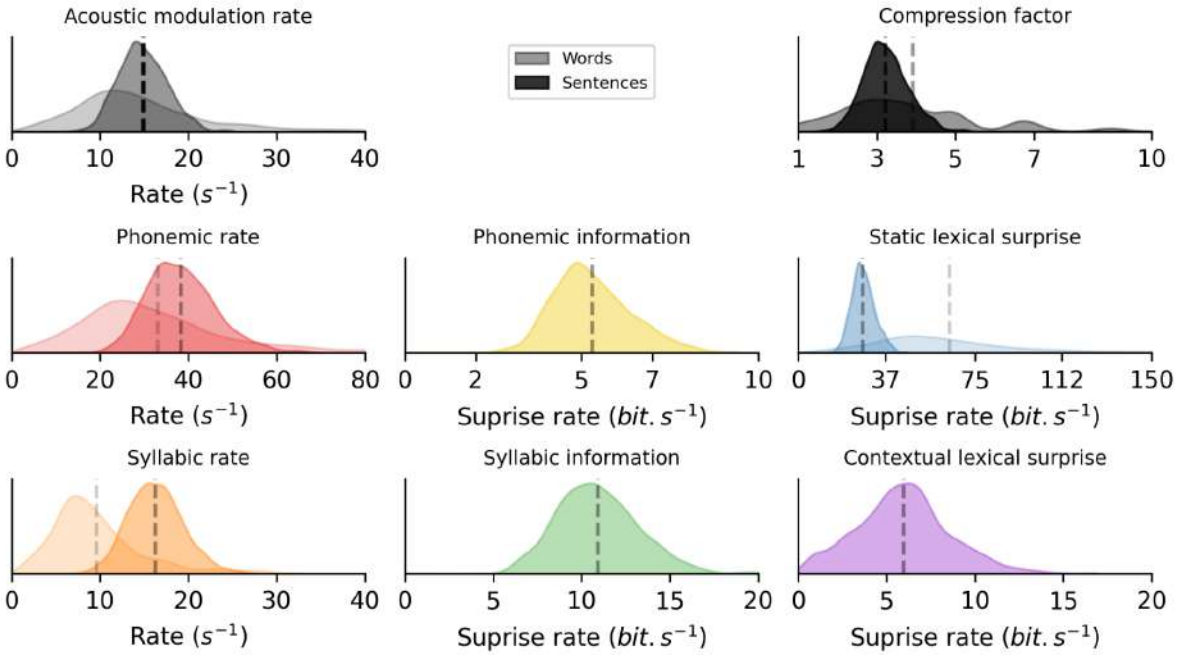


Figure 4. Channel capacity associated with each linguistic feature estimated in experiments 1 (words) and 2 (sentences). *At each trial, the comprehension point – which corresponds to the compression rate at which comprehension emerged – was estimated (upper right panel, see Methods). As each feature significantly impacts comprehension (see Fig. 3), their maximal rate before they begin to negatively impact comprehension can be estimated. Values of each linguistic feature at comprehension points were extracted and aggregated across trials. The resulting distribution provides an estimate of the channel capacity associated with each linguistic feature. Data from experiment 1 (words) is depicted in lighter colors. For each linguistic feature, the channel capacity estimated in experiments 1 and 2 are of the same order of magnitude. Dashed vertical lines indicate the median of each distribution.*

Contextual information rate constrains the flow of natural speech.

We finally estimated whether any linguistic feature was close to its channel capacity in the non-compressed stimulus sets. For each linguistic feature, we thus compared its value at the comprehension point (*i.e.* its channel capacity) and at original speed (*i.e.* its intrinsic statistics) and estimated a percentage of overlap across distributions.

In experiment 2, for each feature, the percentage of overlap between the two distributions was below 1 %, with the exception of the contextual lexical surprise, which was reaching a ~18 % of overlap (a value significantly higher than the others; repeated-measures ANOVA: $F(6,140) = 3482.3$, $p < 0.001$; post-hoc paired t-tests: contextual lexical surprise vs. others: all $p < 0.001$ Tukey-corrected; all other comparisons: $p > 0.9$ Tukey-corrected; Fig. 5, upper right panel). This indicates that it is not unusual in natural speech to observe an amount of contextual lexical surprise close to its channel capacity, while natural speech operates much farther from the channel capacity of the other linguistic features. In experiment 1, the percentage of overlap was around 5% for all features (repeated-measures ANOVA: $F(3,80) = 4.9$, $p = 0.003$; post-hoc paired t-tests, all $p > 0.001$ Tukey-corrected; Fig. Supp. 4).

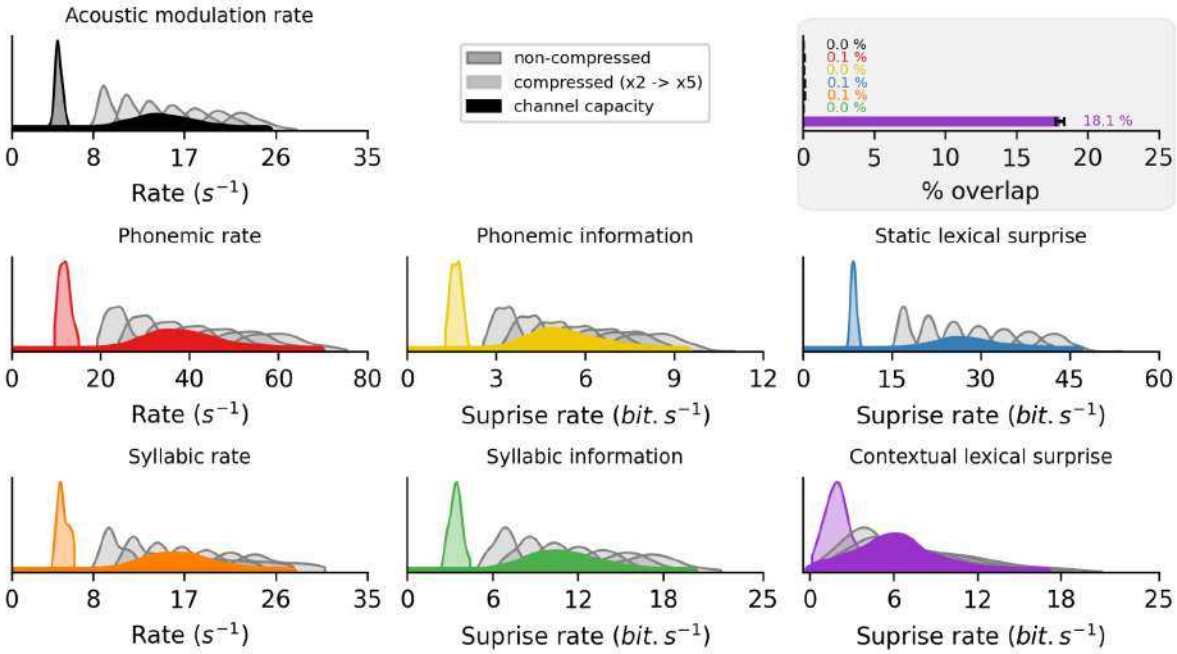


Figure 5. Experiment 2 (sentences). Overlap between the channel capacity associated with each linguistic feature and their generic distribution in the stimulus set. *Distribution of the linguistic features in the selected stimulus set at original speed (non-compressed, lighter color) and at the different compression rates (in grey). Superimposed is their corresponding estimated channel capacity (see Fig. 4; darker color).* ***Upper right (grey panel):*** *Overlap ratio between the channel capacity associated with each linguistic feature and its generic distribution at original speed. Error bars indicate standard error of the mean across participants.*

Discussion

In this study, we investigated the extent to which multilevel linguistic features independently constrain speech comprehension. We expressed each linguistic feature in a number of units per second and derived their associated channel capacity thanks to an innovative experimental paradigm, the compressed speech gating paradigm. Guided by previous lines of research on speech comprehension (Coupé et al., 2019; Ghitza, 2014; Giraud & Poeppel, 2012; Schrimpf et al., 2020), we focused on features encompassing the entire linguistic hierarchy, from acoustic to supra-lexical levels of description, and investigated their individual effect on trial-by-trial performance fluctuations using generalized mixed linear model (GLMM) analyses. We report convergent results using two independent sets of stimuli (words and sentences) and participant sets. Moreover, we showed the robustness of the findings across two different experimental settings (in-lab and online) and complementary analyses (GLMM and LMM). Finally, we reproduce key findings from the literature and report plausible conclusions, compatible with current theoretical models and known biological evidence.

Previous work has focussed on characterizing prominent speech features relevant for comprehension. In particular, speech has been described as an inherently rhythmic phenomenon, in which linguistic information is pseudo-rhythmically transmitted in “packets” (Ghitza, 2014). The theta timescale (4-8 Hz), associated with the main acoustic modulation and the syllabic rates, has been highlighted for its main contribution to speech comprehension (Ahissar et al., 2001; David Poeppel & Assaneo, 2020). Moreover, speech-specific temporal organisation is thought to be reflective of an evolutionary attempt to maximize information transfer given cognitive and neural constraints (Christiansen & Chater, 2016). Accordingly, recent experimental evidence suggests that despite multiple differences, languages are highly similar in terms of average rate of transmission of information (Coupé et al., 2019). Our work is a critical extension of these previous lines of research, by directly comparing multiple relevant features and timescales for speech comprehension into a common measurement framework.

We first behaviorally confirmed human impressive ability to cope with highly speeded speech but also showed a collapse of language comprehension when spoken stimuli presentation rate exceeds a given threshold, i.e. beyond a compression factor of 3 (Dupoux & Green, 1997; Foulke & Sticht, 1969; Ghitza, 2014; Nourski et al., 2009). We show that this phenomenon can be explained as the result of a linear combination of multiple processing bottlenecks along the linguistic hierarchy. Corroborating previous findings, we show that the syllabic rate is the strongest determinant of speech comprehension.

Recent theoretical models propose that speech is sampled in parallel at two timescales, corresponding to the syllabic and phonemic rates (Giraud & Poeppel, 2012). To date, experimental evidence only established that specific brain rhythms in the auditory cortex track the acoustic dynamics during speech perception (Gross et al., 2013; Luo & Poeppel, 2007; Peelle, Gross, & Davis, 2013). Here we directly extended these results at the perceptual level by testing the impacts of the acoustic modulation, syllabic and phonemic rates on comprehension with a tightly orthogonalized setup. Our data reveal that these three features independently constrain speech comprehension. In particular, we found that channel capacities associated with acoustic modulation and syllabic rates were at around 15 Hz while the channel capacity associated with the phonemic rate was at around 35 Hz. These values parallel theoretical considerations and neurophysiological observations (Giraud & Poeppel, 2012; Giroud et al., 2020) and provide a behavioral validation that phonemic sampling

occurs at such a rate (see also (Marchesotti et al., 2020). While the acoustic modulation and syllabic rates are often reduced to one another, they are dissociable (see also (Schmidt et al., 2021), are associated with different processing bottlenecks, but both unfold at around 5 Hz in natural speech and have a channel capacity of around 15 Hz. This result strongly suggests that both low-level acoustic and language-specific rhythmic processes contribute to speech comprehension. The channel capacities estimated for higher-order linguistic features cannot be compared with anything currently known in the literature. These results provide directly testable hypotheses for future human neurophysiology experiments.

Additionally, by developing a normative measurement framework, we bridged speech perception studies with the domains of psycholinguistics, computational linguistics and natural language processing. First, our data reveal a mild adversarial effect of information rate at the phonemic and syllabic scales on speech comprehension. Whether these effects are similar across languages remains an open question. However, previous experimental evidence supports the view that the channel capacities that we estimated would reflect the general human cognitive architecture or the ecological language niche (Coupé et al., 2019; Pellegrino et al., 2011). Second, we show that the respective impact on comprehension of the syllabic rate, the static lexical surprise rate (derived from the lexical frequency) and the contextual lexical surprise rate (derived from a deep neural transformers model) are of the same order of magnitude, but with the syllabic rate having the largest influence.

Among the seven factors investigated in this study, four pertain to information processing in the sense of Shannon's theory of communication. Static and contextual lexical surprises are directly related to the participants' linguistic expectations: both unusual words and sentence structures hinder the capacity to overcome the challenge caused by a high compression rate. Noteworthy is that phonemic and syllabic information rates also have an impact – albeit more limited – on comprehension, in addition to the lexical level. Previous studies highlighting the importance of information rate did not disentangle the syllabic rate from the syllable and lexical information. In the present study, we investigated the syllabic (/phonemic) functional loads, viz. the importance of correctly identifying the presented syllable (/phoneme) to access the target word. In other words, misperceiving a high functional load syllable (/phoneme) may lead to a wrong identification at the word level. Our study thus reveals the role of these phonemic and syllabic contrastive information once the lexical linguistic expectations are taken into account.

We also addressed whether in natural speech and at normal speed, the intrinsic statistics associated with each linguistic feature are already close to their channel capacity. Apart from contextual information, all other features' generic statistics are below their respective channel capacity. Based on those results, we propose that contextual lexical surprise is an important constraint regarding the rate at which natural speech unfolds. Accordingly, speech production and perception can be envisioned as a dynamical information processing cycle, in which the speaker and the listener are two elements in interaction within one closed-loop converging system (Ahissar & Assa, 2016). While in this study we approach the question from the perception side, to delimitate the highest rate at which linguistic inputs can be processed, it would be of great interest to look at the same phenomenon from the production side and determine whether constraints imposed on speech comprehension have some equivalents in speech production. Related to this, investigating whether and which channel capacities can be extended by training could be a powerful way to optimise rehabilitation strategies in patients suffering from speech impairments.

Artificially compressing speech can lead to a degradation of the quality of the linguistic information. This can cause comprehension to drop as linguistic features may most efficiently be represented at their natural rates in the auditory system. However, previous work has repeatedly demonstrated that limitations in compressed speech comprehension are not due to limited capacities in acoustic information encoding. Neural activity recorded in the primary auditory cortex can indeed track the acoustic modulation rate even well outside of the intelligibility range (Nourski et al., 2009; Pefkou et al., 2017). This feat is putatively rendered possible by the short temporal integration windows of early auditory areas (Giroud et al., 2020; Lerner et al., 2014; David Poeppel, 2003). Conversely, the degraded comprehension of speeded speech is thought to arise from limitations of higher order brain areas in their speech-decoding capacities (Vagharchakian et al., 2012). A further argument in favor of this interpretation is that inserting delays between segments of highly compressed speech restores comprehension (Ghitza & Greenberg, 2009), highlighting the fact that is not a problem of stimulus encoding processing but rather a limitation in the time needed to decode the information present in the acoustic signal (Pefkou et al., 2017). By using time-compressed speech, we artificially increased the amount of information per time unit, leading to a drop in comprehension as a result of multilevel limited channel capacities, reflecting internal processes which can not keep up with the overflow of information. This saturation can be considered as analogous to attentional blink and psychological refractory period phenomena (Pashler, 1984; Raymond, Shapiro, & Arnell, 1992; Sigman & Dehaene, 2008) or more complex theoretical frameworks (S Marti, King, & Dehaene, 2015; Sébastien Marti & Dehaene, 2017), which suggests that the complexity of an integration operation defines its channel capacity. Our data are in accordance with this idea, as we showed that multilevel linguistic features predict accelerated speech comprehension performance. One question we can not answer is whether this is the result of a serial chain of processes or of competing parallel processes, or both. Further work using time-resolved measurements of comprehension could adjudicate between these concurrent hypotheses.

Finally, while we used meaningful sentences and words derived from large databases, due to experimental conditions, we artificially accelerated the spoken material to carefully control for speed variations. This controlled experimental task may seem somewhat unnatural but we show that the compressed speech gating paradigm is sensitive to linguistic features that have been shown to influence language processing in more classical experimental settings. Importantly this paradigm allows comparing in a generic framework different linguistic features from previously distinct subfields in the language domain. While the model approach comparison used in this work only affords relative conclusions, it undoubtedly paves the way for more thorough investigations of the effects of multilevel linguistic features on speech comprehension. Thanks to an innovative paradigm and stimuli selection procedure, our approach unifies a diverse literature under the unique concept of channel capacity. Our findings highlight the relevance of using both natural speech material (despite being more methodologically constraining) and a normative measurement framework to study speech comprehension. We hope that this work will settle the ground for further explorations of speech comprehension mechanisms at the interface of multiple linguistic research fields.

Materials and Methods

Participants.

A total of 42 native French speakers (26 females, mean age 23.2 y, standard deviation ± 2.5, range [20, 30]) took part in two behavioral experiments, (21 in experiment 1 and 21 in experiment 2). For the first experiment, participants were recruited from Aix-Marseille University. For the second experiment, participants were recruited online from Aix-Marseille University's student group to perform the experiment through the FindingFive online platform. All participants reported normal audition and no history of neurological or psychiatric disorders. They provided informed consent prior to the experimental session and received financial compensation for their participation (20 euros). The experiments followed the local ethics guidelines from Aix-Marseille University.

Stimuli.

Speech stimuli. The stimuli in experiment 1 consisted of 251 monosyllabic French words drawn from a set of 1,100 monosyllabic words listed in the Lexique database (New et al., 2004). The stimuli in experiment 2 consisted in 100 seven-word-long French sentences drawn from a set of 14,000 seven-word sentences listed in the Web Inventory of Transcribed and Translated Talks database (WTI3, Cettolo et al., 2012). For both experiments, the text stimuli were then synthesized in auditory stimuli using Google Cloud Text-to-Speech (Google, Mountain View, CA, 2020, the female voice, "fr-FR-Wavenet-C"). Stimuli were selected on the basis of their characteristic linguistic features. For that, each stimulus at original speed was characterized by a vector composed of five features in experiment 1 and seven features in experiment 2. These linguistic features characterize the stimuli at different levels of processing, from acoustic to supra-lexical properties. Importantly, each feature was estimated in a number of units per second (i.e., in rate, or bit/s) to allow comparing their respective importance on speech comprehension (Coupé et al., 2019; Pellegrino et al., 2011; Reed & Durlach, 1998). The features were the following:

Acoustic modulation rate: it corresponds to the main acoustic modulation rate present in the speech signal. For each stimulus (words or sentences), the wideband envelope of the speech waveform was estimated (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Z. M. Smith et al., 2002) : the raw speech waveform was band-pass filtered into 32 frequency bands from 80 to 8,500 Hz with a logarithmic spacing, modelling the cochlear frequency decomposition. The absolute value of the Hilbert transform of each band-passed signal was extracted and summed across bands. The resulting envelope time-course was downsampled to 1000 Hz. Then, we used Welch's method (Virtanen et al., 2020) to estimate the power spectral density of the envelope, resulting in a modulation spectrum between 1 and 215 Hz with a 0.1 Hz resolution. This was done for each stimulus. Finally, the center frequency of each spectrum was extracted by taking the global maximum value of each modulation spectrum. The acoustic modulation rate was expressed in Hz.

Phonemic rate: it corresponds to the number of phonemes presented per second. It was computed by dividing the number of phonemes (retrieved from the canonical pronunciation provided in the Lexique database (New, Pallier, Brysbaert, & Ferrand, 2004)) by the duration of the stimulus. The phonemic rate was expressed in Hz.

Syllabic rate: same as the phonemic rate but for syllables. It was also expressed in Hz.

Phonemic information rate: it measures how much information, defined by Shannon's theory of communication, is carried by each phoneme. In order to approach this level from a perspective different from the lexical level described below, we adopted a methodology based on the contrastive role of the phonemes in keeping the words different in the French lexicon. For each distinct phoneme, its contrastive role was computed as its relative functional load (Oh et al., 2015). The functional load allows calculating the relative importance of a phoneme for a given language. More specifically, it quantifies its importance in terms of avoiding homophony keeping the words distinct in the lexicon, given their frequency of usage. The phonemic information rate is consequently defined for each stimulus as the sum of its phonemic functional loads divided by its duration. This feature was estimated from written data derived from the Lexique database. The phonemic information rate was expressed in bits per second.

Syllabic information rate: same as phonemic information rate but for syllables. It was also expressed in bits per second.

Static lexical surprise rate: Derived from the lexical frequency, it measures the unexpectedness of a word without reference to the surrounding context. It was computed as the negative base 2 logarithm of the unconditional probability of a word $-\log_2 P(\text{word})$, where P(word) is the lexical frequency of the word. The lexical frequency was the frequency of occurrence in the Lexique database. In experiment 1, the static lexical surprise was divided by the stimulus duration. In experiment 2, as stimuli were seven-word sentences, the static lexical surprise of each individual word composing the sentences was summed before dividing by the duration of the stimulus. The static lexical surprise was expressed in bits per second.

Contextual lexical surprise rate: Derived from a deep neural transformers model, it measures the unexpectedness of a word given the sentence context. It was computed as the negative base 2 logarithm of the conditional probability of a word $-\log_2 P(\text{word}|\text{context})$, where P(word|context) is the probability of a word estimated by the french Bidirectional Encoder Representations from Transformers CamemBERT (L. Martin et al., 2020). This transformer network is a bidirectional-attention model that uses a series of multi-head attention operations to learn context-sensitive representations for each word in an input sentence in a self-supervised way by predicting a missing word given the surrounding contexts in large text corporas. As the model is bidirectional, it employs both preceding and subsequent context of a word when predicting it. We used the HuggingFace transformers Python package (Wolf et al., 2020) to access the pre-trained CamemBERT model with no further fine-tuning. Each individual sentence stimulus was passed through CamemBERT and the pooled output was averaged over the seven words contained in the sentence. This quantity was finally divided by the stimulus duration. As a context is needed to estimate the contextual lexical surprise, it was only computed for experiment 2, where stimuli are sentences. The contextual lexical surprise was expressed in bits per second.

Procedure and Paradigm.

Orthogonalisation procedure to select the stimulus sets. In order to avoid collinearity issues due to correlations between features across stimuli, we developed a custom-made leave-one out iterative algorithm to select stimuli with low correlation between features. The algorithm starts with the complete original database (1,100 words in experiment 1 and 14,000 sentences in experiment 2) and computes the correlation between each pair of features (5-7 features, 10-21 correlations in total in experiment 1 and 2 respectively). Then, the algorithm performs a leave-one-out procedure: it removes

one stimulus, recomputes the correlation matrix on this reduced set and estimates the specific contribution of the one stimulus on the original correlation matrix, by comparing the correlation matrices of the full and reduced stimuli sets. This processing step is repeated until all items have been removed once. The 10 percent stimuli that led to the most significant increase in correlation across features are discarded. The algorithm then iterates on this newly selected reduced stimuli set. The algorithm stops when the number of stimuli is equal to 251 (words) in experiment 1 and 100 (sentences) in experiment 2. A last check ensured that the correlations between features were all below 0.15.

Representativeness of the selected stimulus sets. The representativeness of the final selected stimulus sets in comparison to the original datasets was assessed for each feature. This was performed to ensure that any theoretical conclusions derived from the results obtained from a limited subset of stimuli could generalize to a larger corpus-based dataset. To do so, we computed the value of the features for the complete datasets, hence providing a relatively good estimate of the ecological distribution of each feature. Two indexes were computed to control that each feature's distribution in the selected stimulus sets was similar to its distribution of the original datasets: i) the ratio between the means, ii) the ratio between the variances. A value close to one for both indexes indicates a good match between the distributions in the original dataset and in the selected stimulus sets. Finally, the correlation matrices between the features in the selected stimulus sets and the features in the original datasets were compared.

Time compression. Time compressed versions of each stimulus were created. The audio waveforms were linearly compressed at rates 1, 2, 2.2, 2.5, 2.9, 3.5, 4.3, 5.6, 8 and 10 of the original recording in experiment 1 and at rates 2, 2.5, 3, 3.5, 4, 4.5 and 5 in experiment 2. A compression rate of 2 indicates that the duration of the time-compressed version of the audio file is equal to half of the natural duration. The compression rates in experiment 2 were adjusted on the basis of the results of experiment 1. The PSOLA algorithm implemented in the Parselmouth Python package based on PRAAT (Boersma, 2001; Jadoul, Thompson, & de Boer, 2018; Moulines & Charpentier, 1990) was used to modify the duration of the audio stimulus without altering the original pitch contour. Audio stimuli were normalized in amplitude and digitized at 44.1 KHz. This resulted in 2510 audio stimuli (251 words x 10 compression rates) in experiment 1 and 700 audio stimuli (100 sentences x 7 compression rates) in experiment 2. A manual check was performed to ensure that the compression procedure did not insert salient quirks.

Paradigm. The two behavioral experiments consisted in a modified version of the gating paradigm (Grosjean, 1980) using time-compressed speech stimuli. Participants were presented with 10 time-compressed versions of isolated words in experiment 1 and 7 time-compressed versions of seven-words sentences in experiment 2. Each trial consisted in the successive presentation of different time compressed versions of the same audio stimulus, in an incremental fashion, starting with the most compressed version of the stimulus (gate n°1) and ending with the least compressed version (either gate n°10 or n°7). After each audio presentation, participants were asked to either type on the keyboard (experiment 1) or repeat in the microphone (experiment 2) what they heard and then to press enter to continue to the next gate. Participants were instructed that each auditory stimulus was meaningful and difficult to understand at the highest compression rates. In order to get familiarized with the task, participants completed three practice trials before the experiment. Each experiment was composed of two sessions of approximately 50 minutes each. The sessions included several breaks for the participants to stay vigilant and focussed throughout the experiment. Each participant was presented with the stimuli in a pseudo-randomized order. The experiment was self-paced and there

were no time constraints. The two sessions were performed at most one week apart. The paradigm used in both experiments incorporated a transcription task which required participants to explicitly recognise, recall, and either reproduce each isolated word or each word of the sentence. It provided a fine-grained accuracy measure associated with focused and extensive linguistic processing.

Experimental setup. Experiment 1 was implemented in Python with the expyriment package (Krause & Lindemann, 2014) and run on a ASUS UX31 laptop. The program presented the audio stimuli binaurally at a comfortable hearing level via headphones (Sennheiser HD 250 linear) and recorded the participants' written responses. Participants came to the laboratory and performed the two sessions in an anechoic room. Due to the Covid-19 outbreak, a different set of participants undertook experiment 2 online via the experimental platform FindingFive (FindingFive, 2019). The procedure was the same except that participants were instructed to record their answers with a microphone (instead of typing them) to optimize the duration of the experiment.

Data analyses.

Data scoring. Speech comprehension was scored 1 if the response was correct (grammatical errors were allowed) and 0 if the response was incorrect or if no answer was given. In experiment 2, participants' audio responses were first transcribed using Google Cloud Speech-to-Text (Google, Mountain View, CA, 2018) and checked manually for mistakes or inconsistencies.

General linear mixed model (GLMM) analysis. Participant's responses (0: incorrect, 1: correct) were analyzed using Generalized Linear Mixed Models (GLMM; (Quené & van den Bergh, 2008) with a logistic link function using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (version 3.5.1,Team, n.d.). The datasets were composed of 52,710 responses in experiment 1 (21 participants x 251 words x 10 compression rates) and 102,900 responses in experiment 2 (21 participants x 100 sentences x 7 words x 7 compression rates). Acoustic modulation rate, phonemic rate, syllabic rate, phonemic information rate and static lexical surprise were entered as fixed effects in experiment 1. Participants and compression rates were entered as random effects. The model was expressed as follows in lme4 syntax:

glmer(performance ~ 1 + scale(phonemic rate) + scale(syllabic rate) + scale(phonemic information rate) + scale(static lexical surprise) + (1 | compression rate) + (1 | participant), family = binomial(link = logit))

In experiment 2, the model was the same except that syllabic information rate and contextual lexical surprise were added as fixed effects. The model was:

glmer(performance ~ 1 + scale(phonemic rate) + scale(syllabic rate) + scale(syllabic information rate) + scale(phonemic information rate) + scale(static lexical surprise) + scale(contextual lexical surprise) + (1 | compression rate) + (1 | participant), family = binomial(link = logit))

No interaction terms were estimated in the models. First, models including all the possible interactions failed to converge. Second, converging models that included a subset of interactions only very marginally increased the percentage of variance explained in the behavioral responses (marginal and conditional R^2). These latter are well and best captured by the main effects.

Post-hoc comparisons between the resulting estimates associated with each feature were conducted using the glht function from the multcomp package in R (Hothorn, Bretz, Westfall, & Heiberger, 2016). All p-values reported were corrected for multiple comparisons using the Holm correction.

Comprehension point determination. For each stimulus, the comprehension point was estimated. It is defined as the compression rate at which participants reached a 75% correct response performance, as predicted by a logistic function. Fitting procedures were performed in R using the glm function from lme4 package (Bates et al., 2015).

Linear mixed model (LMM) analysis. Comprehension points were analyzed using linear mixed models (LMM). This complementary statistical analysis aimed at characterizing the relationship between the values of each feature at normal speed and the comprehension points. The rationale was that if they impact comprehension, the feature values at normal speed are predictors of the compression rate at which comprehension shifts from incorrect to correct. Whereas, in the GLMM analysis, all behavioral responses were entered in the model, the current analysis exploits only the comprehension point in each trial. The final datasets were composed of 5,271 comprehension points in experiment 1 (21 participants x 251 words) and 2,100 comprehension points in experiment 2 (21 participants x 100 sentences). Acoustic rate, phonemic rate, syllabic rate, phonemic information rate and static lexical surprise were entered as fixed effects in experiment 1. Participants and compression rates were entered as random effects. The model was:

lmer(comprehension point ~ 1 + scale(phonemic rate) + scale(syllabic rate) + scale(phonemic information rate) + scale(static lexical surprise) + (1 | participant))

In experiment 2, the model was the same except that syllabic information rate and contextual lexical surprise were added as fixed effects. The model was:

lmer(comprehension point ~ 1 + scale(acoustic modulation rate) + scale(phonemic rate) + scale(syllabic rate) + scale(phonemic information rate) + scale(syllabic information rate) + scale(static lexical surprise) + scale(contextual lexical surprise) + (1 | participant))

Determination of channel capacity associated with each linguistic feature. The processing of each linguistic feature was modeled as a transfer of information through a dedicated channel. Channel capacity is defined as the maximum rate at which information can be transmitted. For each feature, it was estimated using the comprehension point and defined as the value of the feature at the comprehension point.

Overlap between channel capacity and generic features distributions. The overlapping R-package (Pastore, 2018) was used to compute the percentage of overlap between the values of the channel capacity associated with each feature and their generic distribution in the stimulus set at normal speed. The method divides the density distribution into intervals and computes the cumulative sum of minimum values per interval. The result can vary between 0 and 1, where 1 indicates that the two distributions are identical and 0 indicates a complete absence of overlap. The percentage of overlap between feature distributions reveal which feature is already near the upper limit of speech comprehension at normal speed, potentially limiting our ability to cope with higher speed speech.

Model validation. All models were fitted in R (version 3.5.1, (R core, 2020)) and implemented in RStudio (Racine, 2012) using the lme4 package (Bates et al., 2015). Fixed effects were z-transformed to obtain comparable estimates (Schielzeth, 2010). Visual inspection of residual plots was systematically performed to assess deviations from normality or homoscedasticity. Variance inflation factors (VIF) were also checked to ensure that collinearity between fixed effects was absent. Overall, VIF values were generally close to one and no deviations from model assumptions were detected. We tested the significance of the respective full models as compared to the null models by using a

likelihood ratio test (R function anova). Goodness of fit of the models were evaluated and reported using both the marginal and conditional R^2.

Data availability. Numerical data supporting this study will be available on GitHub: https://github.com/DCP-INS/

Code availability. Codes to reproduce the results and figures of this manuscript will be available on GitHub: https://github.com/DCP-INS/

Supplementary Figures

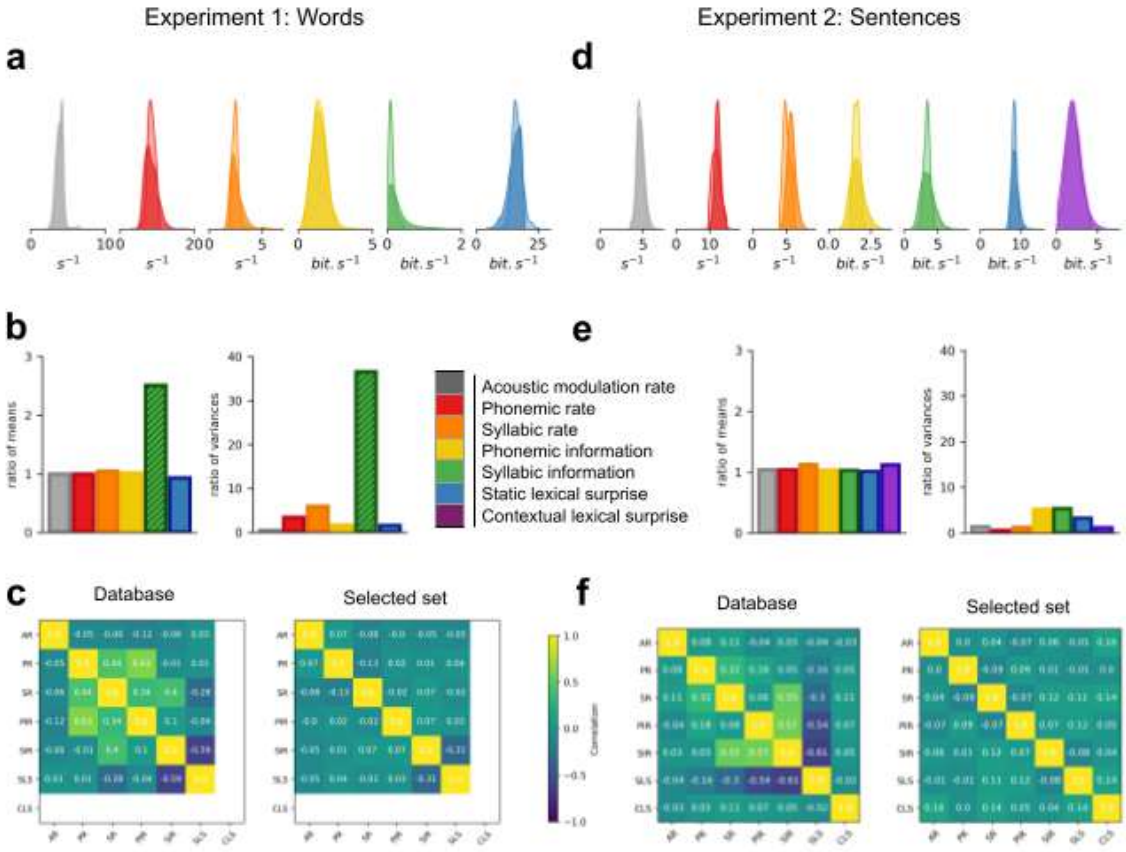


Figure Supplementary 1. Description of the linguistic features in the original database and selected stimulus set, for experiments 1 (a-c) and 2 (d-f). a,d) *Distribution of the linguistic features in the original database (dark colors) and selected stimulus set (light colors), at original speed.* ***b,e)*** *Ratios of means (left) and variance (right) across stimuli, between the selected stimulus set and the database.* ***b)*** *Stripped (green) bars highlight an outlier linguistic feature in experiment 1, for which the selected stimulus set is not representative of the original database.* ***c,f)*** *Correlation matrices between linguistic features in (left) the original database and (right) selected stimulus set. The selection procedure ensured that low correlations (all r < 0.15) across stimuli were present between features in the selected stimulus sets (see Methods). AMR: acoustic modulation rate, PR: phonemic rate, SR: syllabic rate, PIR: phonemic information rate, SIR: syllabic information rate, SLS: static lexical surprise and CLS: contextual lexical surprise.*

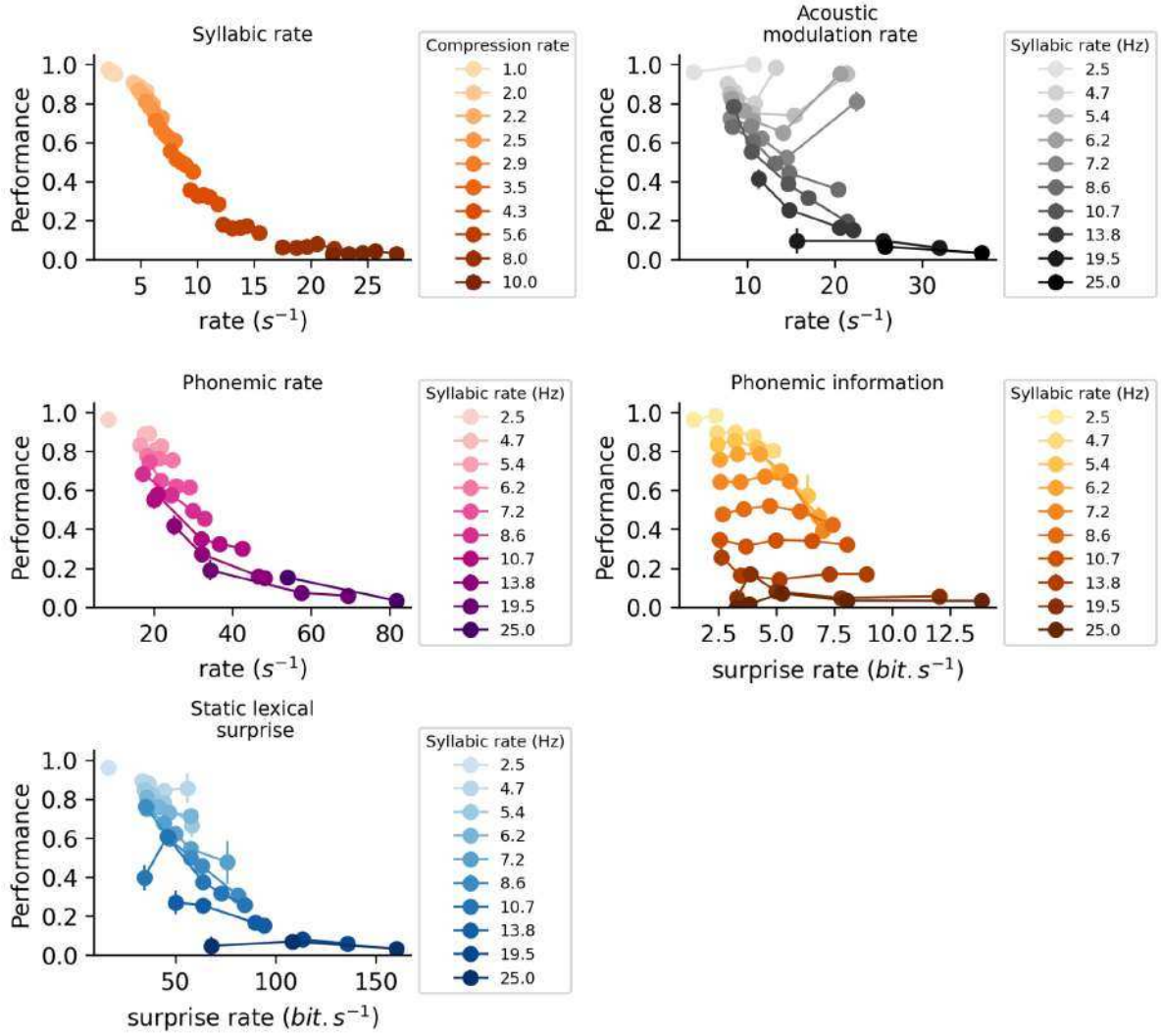


Figure Supplementary 2. Experiment 1. Comprehension performance as a function of the different linguistic features. *Performance is expressed in proportion of correct responses.* ***Upper left panel:*** *Performance sorted as a function of the compression rate (colorscale) and the syllabic rate (y-axis).* ***Other panels:*** *Performance sorted as a function of the syllabic rate (colorscale) and the different linguistic features (y-axes). Data were sorted as a function of the syllabic rate as this feature had the strongest impact on comprehension performance (see Fig. 3) and could thus hide the impact of the other features in this visualisation.*

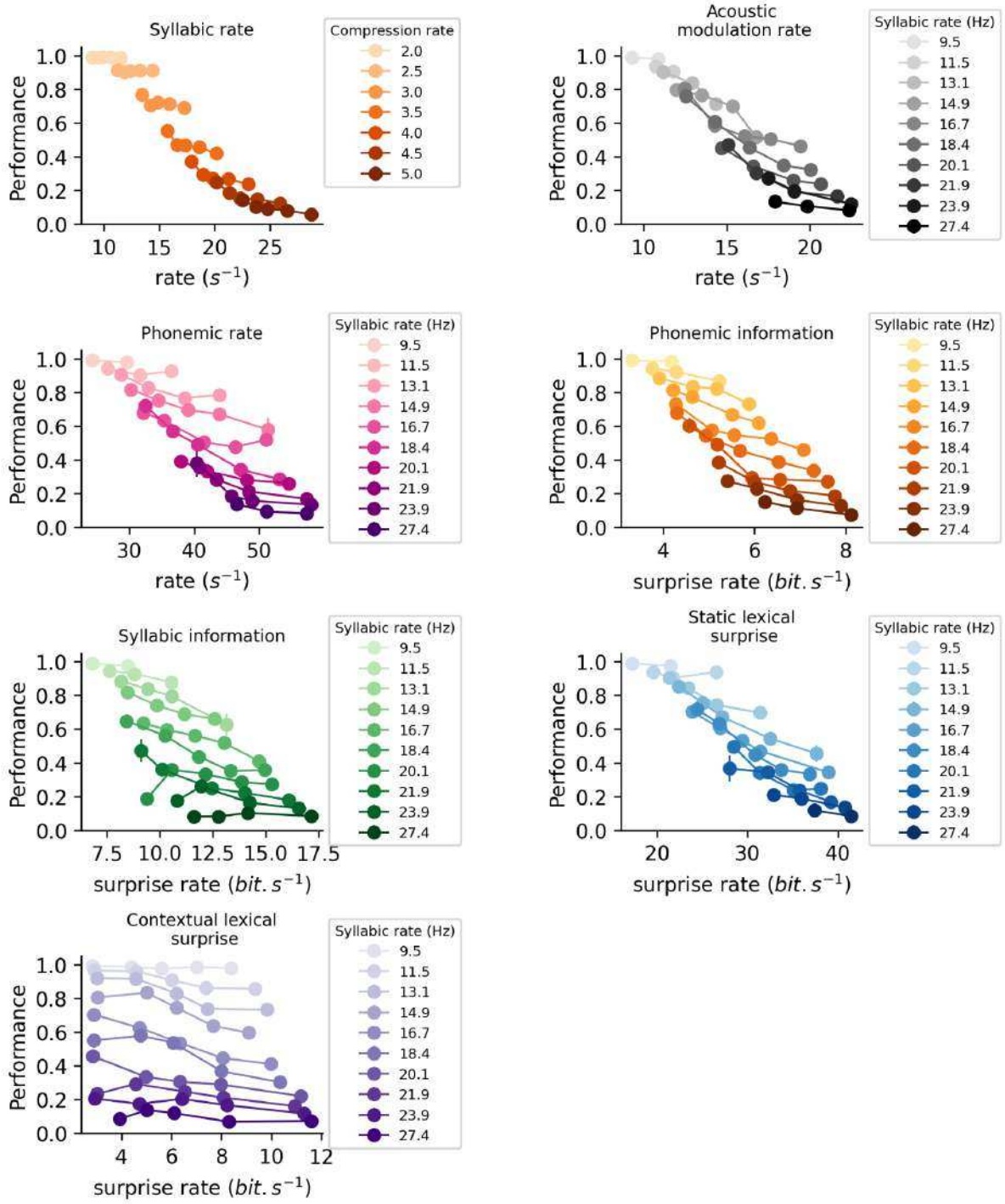


Figure Supplementary 3. Experiment 2. Comprehension performance as a function of the different linguistic features. *Performance is expressed in proportion of correct responses.* ***Upper left panel:*** *Performance sorted as a function of the compression rate (colorscale) and the syllabic rate (y-axis).* ***Other panels:*** *Performance sorted as a function of the syllabic rate (colorscale) and the different linguistic features (y-axes). Data were sorted as a function of the syllabic rate as this feature had the strongest impact on comprehension performance (see Fig. 3) and could thus hide the impact of the other features in this visualisation.*

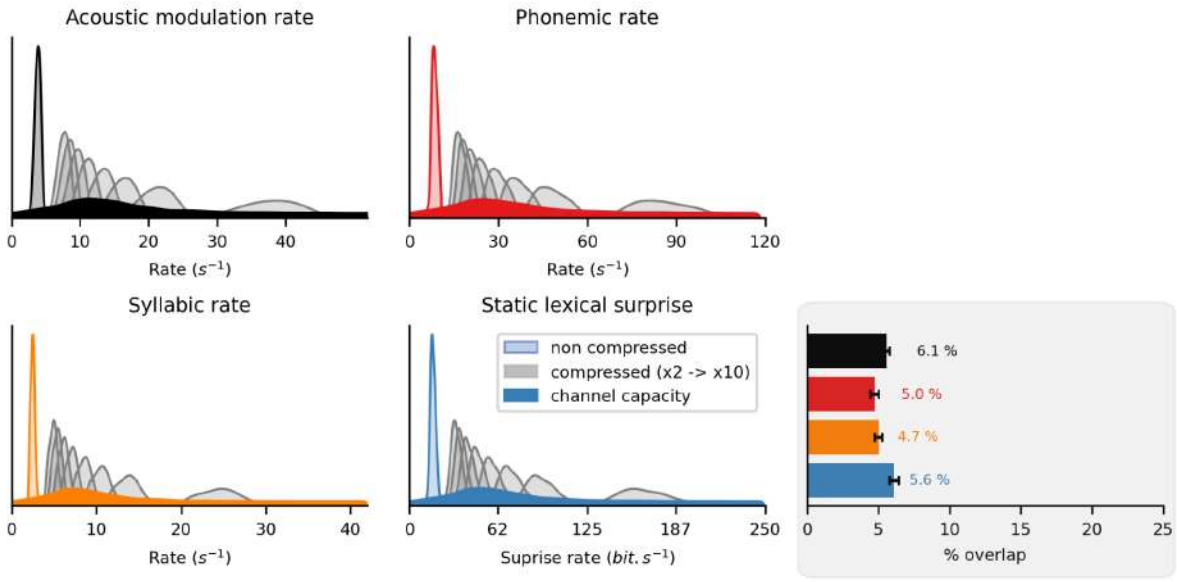


Figure Supplementary 4. Experiment 1 (words). Overlap between the linguistic channel capacities and their generic distribution in the stimulus set. *Distribution of the linguistic features in the selected stimulus set at original speed (non-compressed, lighter color) and at the different compression rates (in grey). Superimposed is the corresponding estimated channel capacity (see Fig. 4; darker color).* ***Lower right (grey panel):*** *Overlap ratio between the channel capacity associated to each linguistic feature and its distribution at original speed. Error bars indicate standard error of the mean across participants.*

Bibliography

Ahissar, E., and Assa, E. (2016). Perception as a closed-loop convergence process. Elife 5.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M.M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci USA 98, 13367–13372.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot. Int.

Brodbeck, C., Hong, L.E., and Simon, J.Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Curr. Biol. 28, 3976-3983.e5.

Brysbaert, M., Lange, M., and Wijnendaele, I.V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. European Journal of Cognitive Psychology 12, 65–85.

Caucheteux, C., Gramfort, A., and King, J.-R. (2021). GPT-2's activations predict the degree of semantic comprehension in the human brain. BioRxiv.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A.A. (2009). The natural statistics of audiovisual speech. PLoS Comput. Biol. 5, e1000436.

Christiansen, M.H., and Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. Behav. Brain Sci. 39, e62.

Coupé, C., Oh, Y., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. Sci. Adv. 5, eaaw2594.

Ding, N., Patel, A.D., Chen, L., Butler, H., Luo, C., and Poeppel, D. (2017). Temporal modulations in speech and music. Neurosci. Biobehav. Rev. 81, 181–187.

Donhauser, P.W., and Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. Neuron 105, 385-393.e9.

Dupoux, E., and Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. Journal of Experimental Psychology: Human Perception and Performance 23, 914–927.

Ferreira, F., Henderson, J.M., Anes, M.D., Weeks, P.A., and McFarlane, D.K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence

from the auditory moving-window technique. Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 324–335.

FindingFive, T. (2019). FindingFive: A web platform for creating, running, and managing your studies in one place (USA: FindingFive Corporation (nonprofit)).

Foulke, E., and Sticht, T.G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. Psychol. Bull. 72, 50–62.

Gagnepain, P., Henson, R.N., and Davis, M.H. (2012). Temporal predictive codes for spoken words in auditory cortex. Curr. Biol. 22, 615–621.

Garvey, W.D. (1953). The intelligibility of speeded speech. J. Exp. Psychol. 45, 102–108.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front. Psychol. 2, 130.

Ghitza, O. (2013). The theta-syllable: a unit of speech information defined by cortical function. Front. Psychol. 4, 138.

Ghitza, O. (2014). Behavioral evidence for the role of cortical θ oscillations in determining auditory channel capacity for speech. Front. Psychol. 5, 652.

Ghitza, O., and Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. Phonetica 66, 113–126.

Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517.

Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., and Morillon, B. (2020). Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. PLoS Biol. 18, e3000207.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., et al. (2020). Thinking ahead: prediction in context as a keystone of language in humans and machines. BioRxiv.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. Percept. Psychophys. 28, 267–283.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. 11, e1001752.

Gwilliams, L., Linzen, T., Poeppel, D., and Marantz, A. (2018). In spoken word recognition, the future predicts the past. J. Neurosci. 38, 7585–7599.

Hasson, U., Yang, E., Vallines, I., Heeger, D.J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. J. Neurosci. 28, 2539–2550.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and de Lange, F.P. (2020). A hierarchy of linguistic predictions during natural language comprehension. BioRxiv.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.

Honey, C.J., Thesen, T., Donner, T.H., Silbert, L.J., Carlson, C.E., Devinsky, O., Doyle, W.K., Rubin, N., Heeger, D.J., and Hasson, U. (2012). Slow cortical dynamics and the accumulation of information over long timescales. Neuron 76, 423–434.

Hothorn, T., Bretz, F., Westfall, P., and Heiberger, R.M. (2016). Package “multcomp.” … Inference in General ….

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

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. J. Phon. 71, 1–15.

Kendall, T. (2013). Speech rate, pause and sociolinguistic variation: studies in corpus sociophonetics.

Krause, F., and Lindemann, O. (2014). Expyriment: a Python library for cognitive and neuroscientific experiments. Behav. Res. Methods 46, 416–428.

Kutas, M., DeLong, K.A., and Smith, N.J. (2011). A Look around at What Lies Ahead: Prediction and Predictability in Language Processing. In Predictions in the Brain, (Oxford University Press), pp. 190–207.

Lerner, Y., Honey, C.J., Silbert, L.J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31, 2906–2915.

Lerner, Y., Honey, C.J., Katkov, M., and Hasson, U. (2014). Temporal scaling of neural responses to compressed and dilated natural speech. J. Neurophysiol. 111, 2433–2444.

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.

Marchesotti, S., Nicolle, J., Merlet, I., Arnal, L.H., Donoghue, J.P., and Giraud, A.-L. (2020). Selective enhancement of low-gamma activity by tACS improves phonemic processing and reading accuracy in dyslexia. PLoS Biol. 18, e3000833.

Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In Proceedings of the

58th Annual Meeting of the Association for Computational Linguistics, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 7203–7219.

Marti, S., and Dehaene, S. (2017). Discrete and continuous mechanisms of temporal selection in rapid visual streams. Nat. Commun. 8, 1955.

Marti, S., King, J.-R., and Dehaene, S. (2015). Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. Neuron 88, 1297–1307.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58, 880–883.

Moulines, E., and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453–467.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: a new French lexical database. Behav. Res. Methods Instrum. Comput. 36, 516–524.

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard, M.A., and Brugge, J.F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. 29, 15564–15574.

Oh, Y.M., Coupé, C., Marsico, E., and Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. J. Phon. 53, 153–176.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. Journal of Experimental Psychology: Human Perception and Performance 10, 358–377.

Pastore, M. (2018). Overlapping: a R package for Estimating Overlapping in Empirical Distributions. JOSS 3, 1023.

Peelle, J.E., and Davis, M.H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. Front. Psychol. 3, 320.

Peelle, J.E., Gross, J., and Davis, M.H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex 23, 1378–1387.

Pefkou, M., Arnal, L.H., Fontolan, L., and Giraud, A.-L. (2017). θ-Band and β-Band Neural Activity Reflects Independent Syllable Tracking and Comprehension of Time-Compressed Speech. J. Neurosci. 37, 7930–7938.

Pellegrino, F., Coupé, C., and Marsico, E. (2011). A CROSS-LANGUAGE PERSPECTIVE ON SPEECH INFORMATION RATE. Language 87, 539–558.

Pickering, M.J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? Trends Cogn Sci (Regul Ed) 11, 105–110.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time.’ Speech Commun. 41, 245–255.

Poeppel, D., and Assaneo, M.F. (2020). Speech rhythms and their neural foundations. Nat. Rev. Neurosci. 21, 322–334.

Quené, H., and van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. J. Mem. Lang. 59, 413–425.

Racine, J.S. (2012). RStudio: A Platform-Independent IDE for R and Sweave. J. Appl. Econ 27, 167–172.

Raymond, J.E., Shapiro, K.L., and Arnell, K.M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? . J. Exp. Psychol. Hum. Percept. Perform. 18, 849–860.

Reed, C.M., and Durlach, N.I. (1998). Note on information transfer rates in human communication. Presence: Teleoperators and Virtual Environments 7, 509–518.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B Biol. Sci. 336, 367–373.

R core, T. (2020). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. Methods Ecol. Evol. 1, 103–113.

Schmidt, F., Chen, Y.-P., Keitel, A., Roesch, S., Hannemann, R., Serman, M., Hauswald, A., and Weisz, N. (2021). Neural speech tracking shifts from the syllabic to the modulation rate of speech as intelligibility decreases. BioRxiv.

Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N.G., Tenenbaum, J.B., and Fedorenko, E. (2020). Artificial neural networks accurately predict language processing in the brain. BioRxiv.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal 27, 379–423.

Sigman, M., and Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. J. Neurosci. 28, 7585–7598.

Smith, Z.M., Delgutte, B., and Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature 416, 87–90.

Sohoglu, E., Peelle, J.E., Carlyon, R.P., and Davis, M.H. (2012). Predictive top-down integration of prior knowledge during speech perception. J. Neurosci. 32, 8443–8453.

Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. 111, 1872–1891.

Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., and Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. J. Neurosci. 32, 9089–9102.

Varnet, L., Ortiz-Barajas, M.C., Erra, R.G., Gervain, J., and Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. J. Acoust. Soc. Am. 142, 1976.

Versfeld, N.J., and Dreschler, W.A. (2002). The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. J. Acoust. Soc. Am. 111, 401–408.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 38–45.

Part III: General discussion

4. Summary of the main findings

The main goal of the experimental works presented in this PhD was to explore and characterize different constraints at play during speech processing. We adopted two complementary views to tackle this issue. The first consisted in the fine-grained characterization of the spatio-temporal neurophysiological activity in auditory cortical regions during sounds and syllables processing. The second approach established the cognitive constraints limiting spoken language comprehension, using rigorously selected experimental material and paradigm.

In the first study, thanks to a large database of intracranial recordings obtained from 96 epileptic patients, we shed light on the neural dynamics in auditory regions during sounds and syllable processing. Given the high spatial and temporal precision offered by stereotactic electroencephalography, we characterized in a meticulous fashion the timescales of auditory information processing at three stages of the cortical auditory hierarchy. Our results demonstrate that the early cortical stages of auditory processing (primary and secondary auditory cortex), when acoustically stimulated with transient pure tone stimuli, respond in a characteristic manner. Such that their activity shows characteristic bimodal spectral profiles. These functional responses are characterized by a main spectral peak in the theta range (around 4–8 Hz), with a secondary peak in the beta/gamma range (around 15–40 Hz). The presence of two concomitant time constants in the dynamics of the evoked response is a strong evidence in favor of the asymmetric sampling in time framework, in which the auditory system makes use of a two-timescale processing mode to perceptually sample acoustic dynamics. Of major importance, we reported no apparent hemispheric difference in these core auditory regions. However, at the level of the associative auditory cortex, our results show the emergence of a strong functional asymmetry, which we observed as a predominant theta/gamma (around 6/40 Hz) activity in left hemisphere and a predominant delta/beta (around 3/15 Hz) activity in right hemisphere. Moreover, we showed that this asymmetrical activity was indicative of language lateralization in 80% of the patients. While most of our results were consistent with previous experimental data and theoretical considerations, they also refined some specific aspects of the conceptual framework. For instance, although a strong functional asymmetry emerged in the higher level of the auditory hierarchy investigated, the spectral profile of the neural activity was not predicted by the asymmetric sampling in time framework. Overall, our study helped characterize with unprecedented details the specific processing timescales of different auditory cortical areas in humans. These general mechanisms are thought to apply to general auditory as well as speech perception.

In the second study, we used concepts from information theory to derive a measure of human capacity to process multi-level information present in the speech signal. Thanks to a carefully crafted stimuli set consisting of either isolated words or natural sentences, characterized at multiple lexical levels (from acoustic to supra lexical), we assessed the

individual effect of each of these levels onto performance. These specific features were investigated based on previous studies in various fields of investigations including linguistics, acoustics, and natural language processing. We developed a modified version of the gating paradigm including multiple presentations of the same stimulus at different degrees of temporal compression. Previous experimental work used time compressed speech and showed that speech comprehension drops when the presentation rate is three times higher than the natural one. In our study we went further as we tightly controlled for the linguistic aspects of each stimulus, thus enabling sound conclusions on their specific influence on the comprehension process. First, we replicated previous results and showed that indeed comprehension dropped beyond a time compression factor of three. We also replicated the now highly robust effect of syllabic rate on speech comprehension and brought evidence that beyond 15 syllables per second comprehension is severely degraded. We also showed that phonemic rate is perceptually relevant and estimates its channel capacity at 35 Hz. These results are consistent with the neurophysiological data from the first study (Giroud et al., 2020) and the asymmetric sampling in time theory/framework (Poeppel, 2003), and indicate that both syllabic and phonemic rates independently shape speech comprehension.

Moreover, our results indicate that while the acoustic modulation rate and the syllabic rate are closely related, both having a frequency of ~5 Hz in natural speech and a channel capacity of ~15 Hz, these two factors are uncorrelated in natural speech and associated to complementary neural processes. Finally, we reveal that when supra-lexical contextual information is present, the impact of all other features is dramatically reduced and that this feature (the unpredictability of words in context) limits the flow of natural speech. Overall, we report robust results, replicated across two different experimental settings (in-lab vs. online), sets of stimuli (words and sentences) and participant sets. We reproduce key findings from the literature, but also go several steps further. Our results are redundant, e.g., validated with complementary analyses. Finally, our conclusions are plausible, compatible with current theoretical models and known biological evidence.

5. Theoretical discussion

5.1 Asymmetry, hemispheric lateralization, why should it matter?

The present PhD work has been conducted at the interface between fundamental research within the Institut de Neurosciences des Systèmes and clinical settings at the epilepsy unit of La Timone hospital. In this particular context, I was faced with different problematics and time scales. Research questions necessitate long phases of development and investigation of mechanisms through careful crafting of experimental paradigm and material. The aim is to get insights into general mechanisms of brain functions supporting language processing and to characterize the underlying neural architecture. By developing integrated models of speech and language processing, the goal of the researchers is to obtain a general view of these processes that can be applied at the level of the general human population. On the other side, clinical questions must be addressed usually on a restricted time scale and information must be extensively gathered during specific periods (for instance during electrode implantation time). Moreover, clinicians are preoccupied with patients' specific cases; concerning the functional asymmetry during language processing, they must determine whether the epileptogenic zone includes critical brain regions for language processing. They must assess the consequences of a chirurgical intervention on the future life conditions and cognitive abilities of the patients on an individual basis. Thus, determining hemispheric language specialisation is of critical importance for the life of the patients.

Because the questions, the immediate needs and consequences are not the same for researchers and clinicians, it constitutes an incredible opportunity to look at the same phenomenon through different lenses and work together hands in hands to bring new and complementary insights on language processing mechanisms. Indeed, pre-surgical settings offer the possibility to record neural activity directly within the human brain. This is a highly valuable opportunity to characterize with unprecedented spatial and temporal resolution brain dynamics at play during speech processing. Such opportunity is at the basis of the first project of this PhD and has allowed to precise previous theoretical models thanks to the granularity offered by the recording modality. Because there is nothing comparable to language in any other species, conclusions related to language processing must be inevitably drawn from human subjects. Epileptic patients due to their conditions also permit the observations of the language system functioning in a very particular state. Because epilepsy is a chronic neurological pathology, it usually leads to anatomical and functional reorganizations. Observing these phenomena help in deriving the limitations of the postulated theoretical language models but also in refining them. So far, the characterization of cognitive functions has benefited equally from observations in healthy and pathological populations. These particular observations in clinical settings can be incorporated into current models and help in getting a better understanding of the determinants of language lateralization. The work resulting from the observation of pathological populations has made it possible to question

classic concepts and has led to the emergence of new theoretical notions. For instance, in contrast to previous beliefs, hemispheric specialization is now conceived as a continuum of two-hemisphere participation and inter-hemispheric cooperation appears to have a major role. Moreover, the functional representation of language is now known to involve not only epicentral regions such as the classically evoked Broca's and Wernicke's areas, but also a large functional network located at the perisylvian level. Language representation presents a strong intra- and inter-individual variability depending on a set of modulating factors of intrinsic and induced nature. Characterizing such factors are of high relevance in the search for better neural markers but also better tools to determine with accuracy anatomical and functional origins of the asymmetry. In the first experimental work presented in this PhD, we have shown that the spectral profile of neural activity within association auditory cortex during passive listening of pure tones was indicative of language lateralization. Currently there is no unequivocal method to predict which is the predominant cerebral hemisphere (inter-hemispheric representation) for language in an individual, nor what is the representation within the predominant hemisphere for language (intra-hemispheric representation). Therefore, an exhaustive evaluation of the anatomical and functional representation of language (at the inter- and intra-hemispheric level) must be performed during the presurgical evaluation. There is currently no specific protocol for this type of assessment. Indeed, each clinical team around the world uses its own protocols for this exploration and most often these protocols are designed by taking language as a monolithic entity and without taking into account behavioral performance. It is important to note that the linguistic operation or process being evaluated can, in the same individual, induce different degrees of hemispheric lateralization. Research work can help in developing paradigms and experimental tasks targeting specific processing stages in order to reveal the predominant hemisphere for language. In the second experimental contribution of this PhD, we developed such an approach and derived carefully designed experimental material from large text corpora. By orthogonalizing the stimuli at multiple linguistic levels from acoustic to supra-lexical, we were able to assess their individual relevance for comprehension. While this work constituted only a first step, ongoing work (not presented in this PhD) seeks to establish the neural substrates supporting each of these specific computations and their role into the functional asymmetry. In addition to their implications for the development of theoretical anatomical and functional models of language comprehension as part of fundamental research, our results have a clinical relevance. Specifically, our behavioral paradigm could be proposed to clinicians as complementary tools to evaluate the risk of post-surgical sequelae such as aphasia. They could also guide in the implementation of cognitive remediation methods. I am deeply convinced that the intricate link between fundamental research and clinical worlds is a key element in the successful discovery of general principles and mechanisms of language comprehension. The study of the functional asymmetry in language is a classic example of such successful cooperation; indeed, since the first clinical observations of Marc Dax, Paul Broca and Carl Wernicke, research and clinical worlds have worked together to get a better understanding of the neural bases of speech and language processing. And thanks to such efforts, today we have a much more detailed account of our capacity to understand each

other. Because this work has tremendous consequences at the theoretical but also highly practical level (language processing is predominantly left lateralized in 90% of the healthy population, and 70% of epilepsy patients; (Tzourio-Mazoyer, Perrone-Bertolotti, Jobard, Mazoyer, & Baciu, 2017), it should be pursued and this is the reason why asymmetry, hemispheric lateralization should definitely matter.

5.2 A more general view on auditory processing

Speech and music represent only a fraction of our everyday auditory environment. Although this previous claim seems trivial, It really hit me during the first covid 19 lockdown. I live in the city center of Marseille which is usually a rich auditory environment made up of lots of artificial noises (car, electric scooters, tram, etc…), more natural ones such as seagulls' vocalizations and speech sounds (lots of happy screaming). However, during this strange period of approximately two months, the acoustic scene was entirely modified. At that time, I remembered wondering what the sound environment in which humans evolved centuries ago might sound like. On a regular basis, I am using a noise cancelling headset while listening to mp3 music or podcasts on my way to the laboratory. This habit also raises questions about the influences such specific auditory experiences (mostly artificial and new on the phylogenetic scale) have on neural processes. Following these considerations, I have been wondering to what extent studying speech perception was an efficient way to uncover general auditory processing principles. I am pretty sure, it is very different thoughts that led Jasmin and colleagues to raise similar questions (Jasmin et al., 2019). In their proposal, they argue for a more domain-general view on auditory processing similar to what has been described in the visual modality. Their starting point is that there are structural differences in terms of anatomy and connectivity across cortical auditory fields. The primary auditory cortex is thinner and contains few feedback connections across cortical layers whereas higher in the auditory cortical hierarchy, the cortex is thicker with more feedback connections linking deeper cortical layers to superficial cortical layers. Moreover, recent experimental data in epileptic patients recorded with electrocorticography, has shown a functional gradient between a caudal and a rostral part of auditory fields during speech processing (Hamilton et al., 2018). The caudal core of the auditory cortex responds very rapidly to sounds onsets and can deal with fast amplitude modulations accurately. In contrast, rostral fields respond more slowly to the starts of sounds and can track slow amplitude envelope modulations with higher accuracy. Based on these anatomical and functional differences, Jasmin and colleagues postulated that caudal and rostral auditory regions may support distinct computations linked to different kinds of perceptual tasks. Their proposal departs from previous conceptual frameworks as it tries to synthesize and articulate results obtained from multiple domain-specific investigations (such as music, speech, etc...) into a wider domain-general auditory explanatory theory. They propose to link the functional characteristics of the caudal regions to sensorimotor and spatial computations, as the temporal response properties of neurons to sound in these regions

make them highly suitable for the guidance of motor responses to sounds in the environment or to self-produced sounds. This is in line with evidence that the recruitment of the motor cortex enhances the accuracy of the tracking of sound events timing (Zalta, Petkoski, & Morillon, 2020). In contrast, the rostral auditory areas are thought to support recognition processes in natural listening conditions, in part due to their neural sensitivities to the structure of sounds over time. Moreover, their higher number of feedback connections across layers may underlie greater processing complexity properties and as such indicate a higher position in the processing hierarchy. This specific configuration could entail the integration of incoming sensory inputs with prior experience which is well suited for speech comprehension for instance. Neural activity in the rostral regions has also demonstrated a sensitivity to contextual and predictive information. Clinical observations in populations of patients with stroke support the caudal and rostral computational distinctiveness as damage to rostral regions impacts sound identification, whereas caudal auditory cortex injury impairs location judgements (Ahveninen et al., 2013; Zündorf, Lewald, & Karnath, 2016). In everyday life conditions, it is highly probable that the mixture of sounds encountered on a daily basis forming multiple auditory objects engage simultaneously caudal and rostral auditory perceptual processes in a particular way which depends on the situation. I have found it desirable to anchor various experimental observations from domain-limited research into a more general explanation of perceptual and neural mechanisms relevant to multiple types of behavior in various natural situations.

5.3 The brain as a dynamical system ?

I work at the Institut de Neurosciences des Systèmes within the Dynamics of Cognitive Process group, as a result my thoughts are obviously influenced by approaches related to dynamical systems. Although I am far from being an expert on the subject, this way of looking at brain function is appealing in many ways. And multiple researchers in various fields related to brain sciences such as in cognitive neurosciences, philosophy but also zoological sciences have called for a shift in the dominant “computationalist” view of looking at the brain function to a dynamical point of view. The dynamical hypothesis claims that cognition is and can be understood as dynamical systems. This view contrasts with the computational approach where cognition is related to information processing with representations at its core principles. One major claim of the dynamical approach is that natural cognitive systems are temporal and continuous in nature.

In his seminal paper, van Gelder described the now famous example of the Watt centrifugal governor to describe the main components of the dynamical hypothesis (Van Gelder, 1995). During the early industrial revolution in England, steam powered engines were used in the industry but many machines required that the source of power be highly uniform. In order to keep the engine speed uniform, adjustments were required in response to changes in boiler pressure and workload. Such challenges necessitated an engineering

solution in the form of a device that could automatically adjust various parameters in an accurate and reactive manner so as to maintain uniform speed; the so-called “governor”. From an engineering perspective, complex problems are often solved by breaking the overall task down into simpler subtasks. Then it is possible to build devices allowing to complete each smaller task. In this specific case the device would have to measure the speed of the engine, compare it to desired speed, measure steam pressure and make adjustments based on the information retrieved to regulate the engine speed in a repetitive manner. The overall task can thus be accurately completed thanks to a serie of successive computations executed by different physical devices such as tachometer, steam pressure meter joined together. Each subtask would have to be carried out repeatedly in the correct order to complete the task correctly. To solve this problem, Watt introduced his centrifugal governor, a cleverly designed device that included a spindle, arms, and balls acting on the throttle valve to regulate the steam (Figure 16).

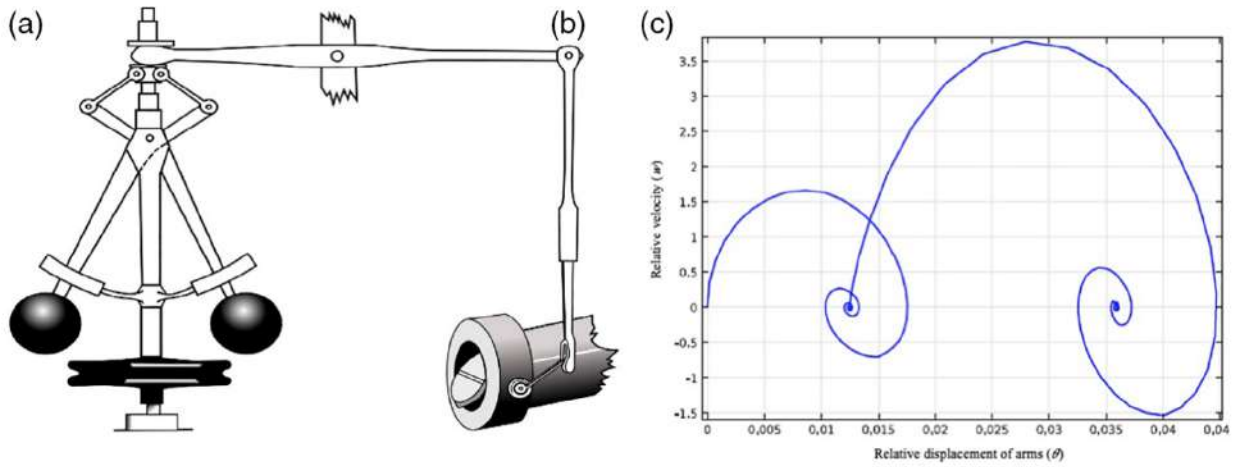


Figure 16 Watt centrifugal governor was developed by James Watt to solve the problem of regulating the speed of a steam engine. Governor comprised of spindle, arms, and balls (a) and throttle valve that regulates steam (b). Phase space plot (c) depicting the two equilibrium points of the system, or, point attractors. Adapted from (Favela, 2020).

Thanks to its specific composition and the gravity, the device did not make use of any measuring devices. Because the device and the throttle valve were coupled in a specific fashion, their specific dynamics allowed a smooth control over power supply. The point of the above presented example is that a complex problem can be solved by a multitude of devices implementing different strategies and bringing various solutions. The Watt governor does not depend on a successive set of computations, it does not depend either on representations. There exist no discrete and identifiable steps in which representations are transformed into different ones. Rather, the system's entire operation is smooth and continuous. The behavior of the system can be described in mathematical terms thanks to differential equations; they

are mathematical functions that describe the evolution of a system over time, where variables are treated as continuous. It is then possible to determine the range of possible values for each variable (state space) by depicting them by means of a phase space plot (Figure 16c).

This approach has been used in the neuroscience community. One famous and fundamental example is the canonical Hodgkin–Huxley model which characterizes the dynamic of action potential derived from data of the squid giant axon (Hodgkin & Huxley, 1952). More recently, the concepts and methods of dynamical systems theory have been used to describe motor dynamics. Motor control is envisaged as dynamic patterns of activity that occur across populations of neurons in order to produce movement. It is in contrast to the view that a part of the cortex gives commands via neurons computing based on representations. In other words, the dynamic systems approach focuses on the dynamics (or rules) of the system that constitutes movement (Churchland et al., 2012).

This thought came from my everyday work in the laboratory, where, in order to either develop paradigms or analyse data, I usually employ the same strategy: I break down problems into small and easy to solve tasks. The result is a single and specific Jupyter Notebook with python code that contains instructions for specific computation for each step to be performed in a specific order. This is pretty efficient I think but not at all natural. It took me some time to get used to it and when I came across van Gelder's paper, his example just echoed in me for a while. In a similar fashion I cannot stop thinking about how the current models of speech processing make use of a specific and precisely ordered set of computations to parse and combine acoustic information to finally lead to meaningful representations. I think that there is a space for complementary views on the question.

Whether the dynamical framework can bring new insights in the field of language processing is still unknown. However, especially in cognitive sciences, where there exists no single theoretical paradigm, explanatory pluralism can be desirable and developing and using methodologies from various fields in an atheoretical manner can foster general principles discoveries but also help refine theories.

5.4 What if neural oscillations aren't real[1] ?

In the last decades, the concept of brain oscillations has emerged as a subject of high interest in both human and animal research (Buzsáki, 2006). In the domain of language and speech processing, neural oscillations are a key element of one of the dominant conceptual frameworks (Giraud & Poeppel, 2012; Peelle & Davis, 2012). This notion has led to an impressive amount of hypothesis and experimental work. While in the very beginning of brain electrophysiology research, neural oscillations were considered to be mere epiphenomena,

[1] In reference to the "Birds aren't real" conspiracy which postulates that birds in the US were killed by the government and then replaced by drones to spy on people. https://www.instagram.com/birdsarentreal/

recent empirical studies argue for a change of perspective on the phenomenon. In the case of speech processing, ongoing debate exists on the nature but also the potential role played by neural oscillations.

First, some researchers have raised the concern that the word oscillations is not entirely appropriate and does not reflect the real nature of the phenomenon. As such it could be misleading and induce the belief that oscillations are pure sine waves whereas there are experimental works in support of the fact that the oscillatory waveform is of non sinusoidal nature (Cole & Voytek, 2017; Doelling & Assaneo, 2021; Donoghue, Schaworonkow, & Voytek, 2021; Quinn et al., 2021). Recently, I came across the term “neural vibrations” in Bergson's book “matière et mémoire” (Bergson, 2020) and I told myself that it would be funny to find such a terminology nowadays in a research paper as a replacement for neural oscillations.

One possible reason for the term “oscillation” comes from the methodological tools used to investigate them. Indeed, the measurements are mostly derived using traditional spectral analysis techniques that assume a sinusoidal basis. And the Fourier decomposition of a signal is a mathematical property, not a biologically based one and may not be the most appropriate means to measure brain activity (Zhou, Melloni, Poeppel, & Ding, 2016). Moreover, contrary to an often (partially) false idea, brain rhythms are less sustained than one would think. This is the result of the standard procedure of averaging data in the spectral domain. Actually, neural oscillations are often observed as transient bursts (Jones, 2016).

Another point of friction concerns the underlying nature of the non-invasively recorded rhythmic activity (Gourévitch, Martin, Postal, & Eggermont, 2020). The relationship between neural entrainment and oscillations remains unclear. At least two different mechanisms have been put forward to explain the non-invasively recorded signals. The first view is based on the assumption that rhythmic stimuli result in the elicitation of a succession of evoked responses that is mistakenly interpreted as a real oscillation (Kojima et al., 2020; Sun, Michalareas, & Poeppel, 2021; Zou et al., 2021). Indeed, usually, when presented with auditory inputs, subcortical and cortical auditory pathways typically elicit an auditory evoked potential to the onset of an acoustic event. Those events are single and highly stereotyped. In the case of rhythmic auditory inputs, multiple auditory evoked potentials could be elicited, and their spectral content contains energy in particular bands that could potentially be confused with oscillations. Another explanation of the phenomenon rests on the phase resetting of the ongoing brain activity. Auditory external stimuli could thus affect concurrently all rhythmic ongoing activities within cortical auditory regions leading to their synchronization and render it visible at the scalp level. Other potential sources of neural oscillations include the interplay between rhythmic components of the auditory inputs and specific nonlinear filtering properties of the auditory regions. Some authors have advanced the hypothesis that the responses of the auditory cortex to external inputs could be reflected by specific resonance frequencies that are characteristic of the area’s time constants linked to perceptual abilities (Zaehle, Lenz, Ohl, & Herrmann, 2010). This view insists on the intrinsic properties of auditory cortical regions that are reflective of architectural and functional constraints and this is the one we adopted on the first project of this PhD thesis (Giroud et al., 2020).

Neural oscillations recorded from the scalp could also reflect the signature of a process. In this view, while not being the proper mechanism resulting in the behavior, neural oscillations would solely indicate that a specific process is at play.

Whether oscillatory activity recorded during auditory processing results from a single or a mixture of the potential mechanisms exposed above is still unknown. In humans, the localization of oscillation generators remains unclear mostly due to the non-invasive nature of the techniques used to investigate them. In this condition, efforts must be done to accurately characterize the neural source(s) underlying the recorded activity.

Experimental work has been conducted to determine whether entrainment is an active or passive mechanism. Multiple lines of evidence suggest that neural oscillations are not only an epiphenomenon or simply a byproduct of thousands of neurons activity. For instance, ten Oever and colleagues, using sub-threshold rhythmic stimulation thus avoiding the elicitation of bottom-up responses, demonstrated the occurrence of the entrainment of neural oscillations before perceptual detection of the auditory stimulus and without apparent sensory evoked responses (Ten Oever et al., 2017).

Figure 17. Sticker stating that "neural oscillations aren't real", found in the canebière, Marseille, France, circa 2020.

A recent study showed that experimental oscillatory neural data in response to music was better explained by an oscillatory model than with an evoked model (Doelling, Assaneo, Bevilacqua, Pesaran, & Poeppel, 2019). This finding supports the evidence of the presence of oscillatory dynamics in the auditory cortex during the processing of rhythmic stimuli. To disentangle proper oscillatory activity from activity resulting from the dynamics of the input stimulus, Kösem and colleagues presented subjects with sentences that changed in speech

rates, while using magnetoencephalographic recordings (Kösem et al., 2018). They demonstrated that neural activity entrained to different speech rhythms and persisted after the change in rate. Moreover, they showed that it had perceptual consequences for determining the identity of subsequent words. Going further, Lizarazu and colleagues demonstrated that the adaptation of the low-frequency oscillations in the auditory regions to different speech rates was dependent on the nesting activity at gamma frequency and its frequency peak changes according to the speech rate (Lizarazu et al., 2019). A recent paradigm shift in understanding brain oscillations with the use of non-invasive brain stimulation has emerged (Vosskuhl, Strüber, & Herrmann, 2018). This line of research ought to reveal brain mechanisms through external electrical or magnetic stimulation and link perceptual outcomes to brain activity in a more direct way. By coupling electroencephalographic recordings and 30 Hz transcranial alternating current stimulation in dyslexic participants, Marchesotti and colleagues selectively and significantly improved phonological processing and reading accuracy in this population (Marchesotti et al., 2020). These results demonstrate a causal role of rhytmic 30 Hz electrical stimulation in phonological processing. Coherent with these results, a functional magnetic resonance imaging study combined with low-frequency transcranial alternating current stimulation indicated that phase entrainment at low-frequency over left lateral temporal regions had a specific and causal influence on neural responses to intelligible speech (Zoefel, Archer-Boyd, & Davis, 2018). Importantly, the effects were absent for unintelligible speech and during sham stimulation.

At the time of writing, the entrainment of neural oscillations question is still vigorously debated. While the field has just lost one of its most influential researchers, neural entrainment is currently considered a fundamental mechanism of brain function, as shown in the latest comprehensive review published by this quiet innovator (Lakatos, Gross, & Thut, 2019; Obleser, 2021). Recently, Doelling and Assaneo, have argued for a step forward in the neural oscillations field (Doelling & Assaneo, 2021). As this phenomenon comprises various heterogeneous neural processes, they suggested that no more effort should be spent on debating the presence of oscillations but rather dig into their specific biophysical mechanisms. They propose as a way forward, the development of quantitative models that make specific predictions, which can be refuted by experimental observations. In addition, it seems reasonable to think that future efforts in experimental design and analytical method development should also contribute to significantly advancing the question. For example, rhythmic stimulation experiments (electric, magnetic, optogenetic, etc...) and simultaneous recordings of field potentials and neuronal spiking will help us overcome major challenges in the field.

6. Limitations and future directions

The present PhD dissertation reports two experiments conceived to explore the various constraints at play during sound and speech processing. As in any experimental project, choices have been made, and methodological as well as experimental considerations have been explored in order to keep the experiments feasible.

First of all, due to the particular nature of the phenomenon under investigation, its mechanisms can only be characterized in humans. Indeed, as already highlighted in this work, language has no real equivalent in any other natural species known so far. Although presurgical evaluation settings are a tremendous opportunity to carefully characterize the processes supporting speech and language understanding, specific concerns apply to all human intracranial electrophysiology studies. The question of the extrapolation of the findings obtained within such a specific population to the general population must be raised. It is known that epilepsy affects the brain organization and can lead to possible specific reorganization even in regions outside of the diagnosed seizure foci. Moreover, the consequences of the use of antiepileptic drugs over extended periods of time on the general brain functioning are not well understood. As a result, evidence obtained in previous works with different recording techniques, or from studies directly comparing data recorded from epileptic patients and healthy participants (Pesnot Lerousseau et al., 2021) are fundamental to corroborate and interpret the intracranial results. By establishing consistency with results of previous studies of auditory cortex, we can be sure that our results are extrapolated to the general population.

Another source of concern comes from the piecemeal sampling of the stereo encephalography recordings. While with most of the non-invasive recording methodologies available it is possible to have a whole head or brain coverage, in the case of intracranial recordings we can only have access to small patches of neural tissues that are moreover dictated by clinical needs. We ensured the reliability of the measurements thanks to a large number of subjects and recording sites. In the first study, we recorded neural response within early cortical auditory processing regions including the posteromedial portion of Heschl's gyrus and the lateral posterior Superior Temporal Gyrus. However, we were not able to record from the lateral portion of Heschl's gyrus and the planum polare. Electrocorticography has a larger surface coverage than stereo-electroencephalography and has been shown to be a good means of characterizing cortical auditory areas under specific conditions (Hamilton et al., 2018, 2021). However, the high temporal and spatial resolution recordings of the caudal part of auditory cortical regions could be highly relevant to study the putative sensorimotor and spatial computational properties of this region (Jasmin et al., 2019).

As previously acknowledged, methodological tools currently used to measure oscillatory activity may not be the most appropriate in this endeavor. Future work should employ recently developed methodologies to remove 1/f trend in neural signals for instance or to take into account the shape of the oscillatory components (Cole & Voytek, 2017; Donoghue et al., 2021, 2020). Alternative options include the use of dictionary learning based techniques

or methods relying on empirical mode decomposition of the neural signal to appropriately model its spectral content (Moreau & Gramfort, 2020). Furthermore, the direct study of the raw signal in the time domain combined with multivariate pattern analysis can be a promising technique to gain insight into auditory cortical processing (Caucheteux et al., 2021).

The use of modulation techniques such as Transcranial Alternating Current Stimulation could definitely participate in the refinement of the theory and knowledge related to auditory and speech processing.

In the second study, we delimited the channel capacity of multiple linguistic features implicated in speech perception. In this experiment we only focused on one of the two sides of communication: perception. Future work should first characterize the constraints at play during speech production. This would provide another angle of analysis on the same issue and reveal the constraints impacting production. It would also allow us to determine whether the same constraints apply to production and perception. Such experimental work could help precise the mechanisms underlying the speech production-perception cycle.

We corroborated previous findings and confirmed that humans can understand speech at a higher rate than usually encountered. However, we used artificially and linearly time-compressed speech as well as text-to-speech engine which may be not the most relevant signal for the brain (Casas et al., 2021). Such limitations can be easily overcome in future work. In our study, this methodological choice was motivated by orthogonalization needs. But in our daily life, we are more and more confronted with artificial speech whether it is in public transportation like trams and subways, or during interactions with our smartphones or connected objects. Although it seems unnatural to us, we have no trouble understanding this type of acoustic signal.

Finally, due to the experimental paradigm adopted in the second study, we were not able to adjudicate between a relatively serial or a parallel chain of processes or whether both were at play during speech perception. Future investigation combining such a paradigm with high temporal resolution neurophysiological recordings could definitely clarify the situation and help reveal the hierarchical or distributed nature of the neural bases of speech perception.

7. Conclusion

During my PhD, I explored two different types of constraints shaping sound and speech processing in humans. The first project focused on the detailed characterization of the functional and architectural constraints of the cortical auditory hierarchy. Based on previous experimental works, we hypothesized to find a functional asymmetry during auditory processing. Thanks to the granularity offered by the stereo encephalography as well as a large number (~100!) of recorded patients and recording sites, we demonstrated a similar bimodal spectral response profile across left and right hemispheres in the first auditory cortical processing stages (the primary and secondary auditory cortex). In response to short acoustic transients, these brain regions specifically demonstrated evoked activity composed of dynamics in the theta (around 4–8 Hz) and beta–gamma (around 15–40 Hz) ranges. In contrast, beyond these first cortical levels of auditory processing, we showed the emergence of an hemispheric asymmetry, with delta and beta band (3/15 Hz) evoked dynamics prevailing in the right hemisphere and theta and gamma band (6/40 Hz) activity prevailing in the left. These results first corroborate previous experimental data and suggest that the pattern of activity in the left association auditory cortex matches the syllabic and phonemic timescales of the natural speech signal, underscoring the importance of a two-timescale processing mode for efficiently analyzing speech. This work shed light on intrinsic neurophysiological constraints shaping auditory processing and permitted refining the current theoretical framework describing the bilateral cortical dynamics involved in auditory and speech processing.

In the second project, we sought to determine the set of cognitive constraints at play during speech processing. We first developed a new experimental paradigm as well as a new experimental corpus composed of naturalistic words and sentences. Capitalizing on the concept of channel capacity from information theory, defined as the maximum rate at which information can be transmitted, we estimated in a fine-grained fashion the influence of seven features, ranging from acoustic to supra-lexical levels of description, on speech comprehension. Our results indicate that all the investigated features are relevant for speech comprehension. Importantly, the syllabic rate had the most impact on comprehension. Our results also suggested that supra-lexical contextual information is the main determinant of the flow of natural speech. Thanks to the normative framework introduced in this work we were able to bridge different conceptual views and quantitatively compare them to get a more general picture of the cognitive determinants of speech understanding.

Overall, our results suggest that spoken language comprehension is shaped by multiple constraints that can be observed at both the neurophysiological and cognitive levels of analyses. In line with recent proposals, our data highlight the major role of a dual timescale for speech processing. Our newly developed paradigm offers a new avenue to better characterize the different processing steps underlying speech comprehension and reveal their respective neural underpinning. It could moreover be used in clinical settings to determine language lateralization in a non-monolithic way and thus reduce postoperative risks of developing language impairments.

Bibliography

Ahissar, E., & Assa, E. (2016). Perception as a closed-loop convergence process. *eLife*, *5*.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–13372.

Ahveninen, J., Huang, S., Nummenmaa, A., Belliveau, J. W., Hung, A.-Y., Jääskeläinen, I. P., Rauschecker, J. P., et al. (2013). Evidence for distinct human auditory cortex regions for sound location versus identity processing. *Nature Communications*, *4*, 2585.

Ainsworth, M., Lee, S., Cunningham, M. O., Traub, R. D., Kopell, N. J., & Whittington, M. A. (2012). Rates and rhythms: a synergistic view of frequency and temporal coding in neuronal networks. *Neuron*, *75*(4), 572–583.

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047.

Anderson, B., Southern, B. D., & Powers, R. E. (1999). Anatomic asymmetries of the posterior superior temporal lobes: a postmortem study. *Neuropsychiatry, neuropsychology, and behavioral neurology*, *12*(4), 247–254.

Annett, M. (1998). Handedness and cerebral dominance: the right shift theory. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *10*(4), 459–469.

Arnal, L. H., Doelling, K. B., & Poeppel, D. (2015). Delta-Beta Coupled Oscillations Underlie Temporal Prediction Accuracy. *Cerebral Cortex*, *25*(9), 3077–3085.

Arnal, L. H., Poeppel, D., & Giraud, A.-L. (2016). A neurophysiological perspective on speech processing in “the neurobiology of language.” *Neurobiology of Language* (pp. 463–478).

Elsevier.

Assaneo, M. F., & Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances*, *4*(2), eaao3842.

Aubanel, V., & Schwartz, J.-L. (2020). The role of isochrony in speech perception in noise. *Scientific Reports*, *10*(1), 19580.

Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, *20*(3), 327–339.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology. Human Perception and Performance*, *10*(3), 340–357.

Barczak, A., O'Connell, M. N., McGinnis, T., Ross, D., Mowery, T., Falchier, A., & Lakatos, P. (2018). Top-down, contextual entrainment of neuronal oscillations in the auditory thalamocortical circuit. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(32), E7605–E7614.

Barth, J. M., Boles, D. B., Giattina, A. A., & Penn, C. E. (2012). Preschool child and adult lateralisation and performance in emotion and language tasks. *Laterality: Asymmetries of Body, Brain and Cognition*, 1–16.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., et al. (2009). Language is a complex adaptive system: position paper. *Language learning*, *59*, 1–26.

Berger, H. (1929). Über das elektroenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*.

Bergson, H. (2020). Matière et mémoire.

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, *29*(3), 191–211.

Bethmann, A., Tempelmann, C., De Bleser, R., Scheich, H., & Brechmann, A. (2007). Determining language laterality by fMRI and dichotic listening. *Brain Research*, *1133*(1), 145–157.

Bisazza, A., Rogers, L. J., & Vallortigara, G. (1998). The origins of cerebral asymmetry: a review of evidence of behavioural and brain lateralization in fishes, reptiles and amphibians. *Neuroscience and Biobehavioral Reviews*, *22*(3), 411–426.

Blank, H., Spangenberg, M., & Davis, M. H. (2018). Neural prediction errors distinguish perception and misperception of speech. *The Journal of Neuroscience*, *38*(27), 6076–6089.

Bless, J. J., Westerhausen, R., von Koss Torkildsen, J., Gudmundsen, M., Kompus, K., & Hugdahl, K. (2015). Laterality across languages: Results from a global dichotic listening study using a smartphone application. *Laterality*, *20*(4), 434–452.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*

Boles, D. B., Barth, J. M., & Merrill, E. C. (2008). Asymmetry and performance: toward a neurodevelopmental theory. *Brain and Cognition*, *66*(2), 124–139.

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, perception & psychophysics*, *79*(1), 333–343.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, *28*(24),

3976-3983.e5.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica*, *49*(3–4), 155–180.

Brysbaert, M., Lange, M., & Wijnendaele, I. V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*(1), 65–85.

Buchsbaum, B. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive science*, *25*(5), 663–678.

Buchsbaum, B. R., Olsen, R. K., Koch, P., & Berman, K. F. (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron*, *48*(4), 687–697.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*(5679), 1926–1929.

Buzsáki, G., Logothetis, N., & Singer, W. (2013). Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. *Neuron*, *80*(3), 751–764.

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press.

Casas, A. S. H., Lajnef, T., Pascarella, A., Guiraud, H., Laaksonen, H., Bayle, D., Jerbi, K., et al. (2021). Neural oscillations track natural but not artificial fast speech: Novel insights from speech-brain coupling using MEG. *Neuroimage*, 118577.

Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*(3), 348–369.

Catani, M., Allin, M. P. G., Husain, M., Pugliese, L., Mesulam, M. M., Murray, R. M., & Jones, D. K. (2007). Symmetries in human brain language pathways correlate with verbal recall. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(43), 17163–17168.

Catani, M., Jones, D. K., & ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, *57*(1), 8–16.

Caucheteux, C., Gramfort, A., & King, J. R. (2021). GPT-2's activations predict the degree of semantic comprehension in the human brain. *BioRxiv*.

Chait, M., Greenberg, S., Arai, T., Simon, J. Z., & Poeppel, D. (2015). Multi-time resolution analysis of speech: evidence from psychophysics. *Frontiers in Neuroscience*, *9*, 214.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432.

Charles-Luce, J., Luce, P. A., & Cluff, M. S. (1990). Retroactive influence of syllable neighborhoods.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975.

Chiarello, C., Welcome, S. E., Halderman, L. K., & Leonard, C. M. (2009). Does degree of asymmetry relate to performance? An investigation of word recognition and reading in consistent and mixed handers. *Brain and Cognition*, *69*(3), 521–530.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, *106*(5), 2719–2732.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, *487*(7405), 51–56.

Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., & Pesaran, B. (2014). Sensory-motor transformations for speech occur bilaterally. *Nature*, *507*(7490), 94–98.

Cole, S. R., & Voytek, B. (2017). Brain oscillations and the importance of waveform shape. *Trends in Cognitive Sciences*, *21*(2), 137–149.

Comrie, B. (1989). Language universals and linguistic typology: Syntax and morphology. *Language universals and linguistic typology: Syntax and morphology*.

Coupé, C., Oh, Y., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Sci. Adv.*, *5*(9), eaaw2594.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147.

Davis, M. H., MacGregor, L. J., Blank, H., Engel, S. A., Helbling, S., Sohoglu, E., & Waschke, L. (2018). Neural oscillations track changes in speech rate shown by MEG adaptation and perceptual after-effects. Presented at the Poster presented at 10th Annual Meeting of the Society for the Neurobiology of Language (SNL).

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, *88*(1), 2–19.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670.

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

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, *81*(Pt B), 181–187.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, *85 Pt 2*, 761–768.

Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(20), 10113–10121.

Doelling, K. B., & Assaneo, M. F. (2021). Neural oscillations are a start toward understanding brain activity rather than the end. *PLoS Biology*, *19*(5), e3001234.

Dolata, J. K., Davis, B. L., & Macneilage, P. F. (2008). Characteristics of the rhythmic organization of vocal babbling: implications for an amodal linguistic rhythm. *Infant Behavior & Development*, *31*(3), 422–431.

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech

processing. *Neuron*, *105*(2), 385-393.e9.

Donoghue, T., Haller, M., Peterson, E. J., Varma, P., Sebastian, P., Gao, R., Noto, T., et al. (2020). Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, *23*(12), 1655–1665.

Donoghue, T., Schaworonkow, N., & Voytek, B. (2021). Methodological considerations for studying neural oscillations. *The European Journal of Neuroscience*.

Dorsaint-Pierre, R., Penhune, V. B., Watkins, K. E., Neelin, P., Lerch, J. P., Bouffard, M., & Zatorre, R. (2006). Asymmetries of the planum temporale and Heschl's gyrus: relationship to language lateralization. *Brain: A Journal of Neurology*, *129*(Pt 5), 1164–1176.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience*, *22*, 567–631.

Driscoll, C., Kei, J., & McPherson, B. (2002). Handedness effects on transient evoked otoacoustic emissions in schoolchildren. *Journal of the American Academy of Audiology*, *13*(08), 403–406.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 914–927.

Du, Y., & Zatorre, R. (2017). Musical training sharpens and bonds ears and tongue to hear speech better. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(51), 13579–13584.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, *5*(3), e1000302.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and linguistics compass*, *6*(10), 635–653.

Esteves, M., Lopes, S. S., Almeida, A., Sousa, N., & Leite-Almeida, H. (2020). Unmasking the relevance of hemispheric asymmetries-Break on through (to the other side). *Progress in Neurobiology*, *192*, 101823.

van Ettinger-Veenstra, H. M., Ragnehed, M., Hällgren, M., Karlsson, T., Landtblom, A. M., Lundberg, P., & Engström, M. (2010). Right-hemispheric brain activation correlates to language performance. *Neuroimage*, *49*(4), 3481–3488.

Everts, R., Lidzba, K., Wilke, M., Kiefer, C., Mordasini, M., Schroth, G., Perrig, W., et al. (2009). Strengthening of laterality of verbal and visuospatial functions during childhood and adolescence. *Human Brain Mapping*, *30*(2), 473–483.

Fant, G. (1971). *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. DE GRUYTER.

Favela, L. H. (2020). Dynamical systems theory in cognitive science and neuroscience. *Philosophy compass*, *15*(8).

Ferreira, F., Henderson, J. M., Anes, M. D., Weeks, P. A., & McFarlane, D. K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 324–335.

Ferrer-i-Cancho, R. (2016). Compression and the origins of Zipf's law for word frequencies. *Complexity*, *21*(S2), 409–411.

FindingFive, T. (2019). *FindingFive: A web platform for creating, running, and managing your studies in one place*. Computer software, USA: FindingFive Corporation (nonprofit).

Finkl, T., Hahne, A., Friederici, A. D., Gerber, J., Mürbe, D., & Anwander, A. (2020). Language without speech: segregating distinct circuits in the human brain. *Cerebral Cortex*, *30*(2), 812–823.

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: unifying

approaches from cognitive neuroscience and comparative cognition. *Physics of life reviews*, *11*(3), 329–364.

Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature Human Behavior*.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). “Who” is saying “what”? Brain-based decoding of human voice and speech. *Science*, *322*(5903), 970–973.

Foulke, E., & Sticht, T. G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, *72*(1), 50–62.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Elsevier BV*.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the annual meeting of the …*.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Friederici, A. D. (2009). Pathways to language: fiber tracts in the human brain. *Trends in Cognitive Sciences*, *13*(4), 175–181.

Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. *Neuron*, *88*(1), 220–235.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615–621.

Galuske, R. A., Schlote, W., Bratzke, H., & Singer, W. (2000). Interhemispheric asymmetries of the modular structure in human temporal cortex. *Science*, *289*(5486), 1946–1949.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125.

Garvey, W. D. (1953). The intelligibility of speeded speech. *Journal of experimental psychology*, *45*(2), 102–108.

Gervain, J., & Geffen, M. N. (2019). Efficient neural coding in auditory and speech perception. *Trends in Neurosciences*, *42*(1), 56–65.

Geschwind, N., & Galaburda, A. M. (1985). Cerebral lateralization. Biological mechanisms, associations, and pathology: I. A hypothesis and a program for research. *Archives of Neurology*, *42*(5), 428–459.

Geschwind, N., & Levitsky, W. (1968). Human brain: left-right asymmetries in temporal speech region. *Science*, *161*(3837), 186–187.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1–2), 113–126.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*, 130.

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Frontiers in Psychology*, *3*, 238.

Ghitza, O. (2013). The theta-syllable: a unit of speech information defined by cortical function. *Frontiers in Psychology*, *4*, 138.

Ghitza, O. (2014). Behavioral evidence for the role of cortical θ oscillations in determining auditory channel capacity for speech. *Frontiers in Psychology*, *5*, 652.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech

perception and production. *Neuron*, *56*(6), 1127–1134.

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

Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., & Morillon, B. (2020). Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. *PLoS Biology*, *18*(3), e3000207.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of phonetics*, *31*(3–4), 305–320.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., et al. (2020). Thinking ahead: prediction in context as a keystone of language in humans and machines. *BioRxiv*.

Gourévitch, B., Martin, C., Postal, O., & Eggermont, J. J. (2020). Oscillations in the auditory system and their possible role. *Neuroscience and Biobehavioral Reviews*, *113*, 507–528.

Greenberg, S., & Ainsworth, W. A. (2004). Speech processing in the auditory system: an overview. *Speech processing in the auditory system* (pp. 1–62). New York: Springer-Verlag.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of phonetics*, *31*(3–4), 465–485.

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech communication*, *29*(2–4), 159–176.

Greenberg, S. (2001). What are the essential cues for understanding spoken language? *The Journal of the Acoustical Society of America*, *109*(5), 2382–2382.

Griffiths, T. D., Uppenkamp, S., Johnsrude, I., Josephs, O., & Patterson, R. D. (2001). Encoding of the temporal regularity of sound in the human brainstem. *Nature*

Neuroscience, *4*(6), 633–637.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Güntürkün, O, Diekamp, B., Manns, M., Nottelmann, F., Prior, H., Schwarz, A., & Skiba, M. (2000). Asymmetry pays: visual lateralization improves discrimination success in pigeons. *Current Biology*, *10*(17), 1079–1081.

Güntürkün, Onur, Ströckens, F., & Ocklenburg, S. (2020). Brain lateralization: A comparative perspective. *Physiological Reviews*, *100*(3), 1019–1063.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *The Journal of Neuroscience*, *38*(35), 7585–7599.

Hall, D. A., Hart, H. C., & Johnsrude, I. S. (2003). Relationships between human auditory cortical structure and function. *Audiology & Neuro-Otology*, *8*(1), 1–18.

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, *28*(12), 1860-1871.e4.

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

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, *184*(18), 4626-4639.e13.

Hartwigsen, G., Bengio, Y., & Bzdok, D. (2021). How does hemispheric specialization

contribute to human-defining cognition? *Neuron*, *109*(13), 2075–2090.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience*, *28*(10), 2539–2550.

Hayes, T. L., & Lewis, D. A. (1996). Magnopyramidal neurons in the anterior motor speech region. Dendritic features and interhemispheric comparisons. *Archives of Neurology*, *53*(12), 1277–1283.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2020). A hierarchy of linguistic predictions during natural language comprehension. *BioRxiv*.

Heschl, R. L. (1878). Ueber die vordere quere Schläfenwindung des menschlichen Grosshirns: aus Anlass der 25-jährigen Jubiläums-Feier der Wiener Landes-Irrenanstalt.

Hickok, G, & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131–138.

Hickok, Gregory, & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402.

Hickok, Gregory. (2012). Computational neuroanatomy of speech production. *Nature Reviews. Neuroscience*, *13*(2), 135–145.

Hirnstein, M., Hugdahl, K., & Hausmann, M. (2014). How brain asymmetry relates to performance - a large-scale dichotic listening study. *Frontiers in Psychology*, *4*, 997.

Hockett, C. F. (1955). A manual of phonology. *A manual of phonology*.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 89–96.

Hockett, C. F. (1966). THE QUANTIFICATION OF FUNCTIONAL LOAD--A LINGUISTIC PROBLEM.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, *117*(4), 500–544.

Holdgraf, C. R., de Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., Knight, R. T., et al. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, *7*, 13654.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, perception & psychophysics*, *72*(5), 1218–1227.

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., et al. (2012). Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, *76*(2), 423–434.

Hothorn, T., Bretz, F., Westfall, P., & Heiberger, R. M. (2016). Package “multcomp.” *… inference in general …*.

Hovsepyan, S., Olasagasti, I., & Giraud, A.-L. (2020). Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature Communications*, *11*(1), 3117.

Howes, D., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of experimental psychology*, *41*(6), 401–410.

Howes, D. (1957). On the Relation between the Intelligibility and Frequency of Occurrence of English Words. *The Journal of the Acoustical Society of America*, *29*(2), 296–305.

Hudspeth, A. J. (2008). Making an effort to listen: mechanical amplification in the ear. *Neuron*, *59*(4), 530–545.

Hugdahl, K., Carlsson, G., Uvebrant, P., & Lundervold, A. J. (1997). Dichotic-listening performance and intracarotid injections of amobarbital in children and adolescents. Preoperative and postoperative comparisons. *Archives of Neurology*, *54*(12), 1494–1500.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience*, *36*(6), 2014–2026.

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

Hutsler, J., & Galuske, R. A. W. (2003). Hemispheric asymmetries in cerebral cortical networks. *Trends in Neurosciences*, *26*(8), 429–435.

Hutsler, Jeffrey J. (2003). The specialized structure of human language cortex: pyramidal cell size asymmetries within auditory and language-associated regions of the temporal lobes. *Brain and Language*, *86*(2), 226–242.

Hutsler, J J, & Gazzaniga, M. S. (1996). Acetylcholinesterase staining in human auditory and language cortices: regional variation of structural features. *Cerebral Cortex*, *6*(2), 260–270.

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

Iacoboni, M. (2008). The role of premotor cortex in speech perception: evidence from fMRI and rTMS. *Journal of physiology, Paris*, *102*(1–3), 31–34.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(6), 1939–1956.

Inbar, M., Grossman, E., & Landau, A. N. (2020). Sequences of Intonation Units form a ~ 1 Hz rhythm. *Scientific Reports*, *10*(1), 15846.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, *10*(11), 480–486.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of phonetics*, *71*, 1–15.

Jamison, H. L., Watkins, K. E., Bishop, D. V. M., & Matthews, P. M. (2006). Hemispheric specialization for processing auditory nonspeech stimuli. *Cerebral Cortex*, *16*(9), 1266–1275.

Jasmin, K., Lima, C. F., & Scott, S. K. (2019). Understanding rostral-caudal auditory cortex contributions to auditory perception. *Nature Reviews. Neuroscience*, *20*(7), 425–434.

Johnson, E. L., Kam, J. W. Y., Tzovara, A., & Knight, R. T. (2020). Insights into human cognition from intracranial EEG: A review of audition, memory, internal cognition, and causality. *Journal of Neural Engineering*, *17*(5), 051001.

Jones, S. R. (2016). When brain rhythms aren't "rhythmic": implication for their mechanisms and meaning. *Current Opinion in Neurobiology*, *40*, 72–80.

Joos, M. (1948). Acoustic Phonetics. *Language*, *24*(2), 5.

Kannan, P. M., & Lipscomb, D. M. (1974). Letter: Bilateral hearing asymmetry in a large population. *The Journal of the Acoustical Society of America*, *55*(5), 1092–1094.

Kaufeld, G., Bosker, H. R., Ten Oever, S., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic Structure and Meaning Organize Neural Oscillations into a Content-Specific Hierarchy. *The Journal of Neuroscience*, *40*(49), 9467–9475.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, *16*(3), e2004473.

Keitel, A., & Gross, J. (2016). Individual Human Brain Areas Can Be Identified from Their Characteristic Spectral Activation Fingerprints. *PLoS Biology*, *14*(6), e1002498.

Keitel, A., Ince, R. A. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *Neuroimage*, *147*, 32–42.

Kendall, T. (2013). Speech rate, pause and sociolinguistic variation: studies in corpus sociophonetics.

Khalighinejad, B., Patel, P., Herrero, J. L., Bickel, S., Mehta, A. D., & Mesgarani, N. (2021). Functional characterization of human Heschl's gyrus in response to natural speech. *Neuroimage*, *235*, 118003.

Khoshkhoo, S., Leonard, M. K., Mesgarani, N., & Chang, E. F. (2018). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. *Brain and Language*, *187*, 83–91.

Kimura, D. (1961). Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *15*(3), 166–171.

King, F., & Kimura, D. (1972). Left-ear superiority in dichotic perception of vocal nonverbal sounds. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *26*(2), 111–116.

King, S., & Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer speech & language*, *14*(4), 333–353.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, *67*(3), 971–995.

Kojima, K., Oganian, Y., Cai, C., Findlay, A., Chang, E. F., & Nagarajan, S. (2020). Low-frequency neural tracking of speech envelope reflects evoked responses to acoustic edges. *BioRxiv*.

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A. S., Jensen, O., & Hagoort, P. (2018). Neural entrainment determines the words we hear. *Current Biology*, *28*(18), 2867-2875.e3.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017).

Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, *93*(3), 480–490.

Krause, F., & Lindemann, O. (2014). Expyriment: a Python library for cognitive and neuroscientific experiments. *Behavior Research Methods*, *46*(2), 416–428.

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J.-F. L., & Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(31), 11238–11245.

Künzel, H. J. (2013). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech Language and the Law*, *4*(1), 48–83.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A Look around at What Lies Ahead: Prediction and Predictability in Language Processing. *Predictions in the brain* (pp. 190–207). Oxford University Press.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

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

Lakatos, P, Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*(2), 279–292.

Lakatos, Peter, Gross, J., & Thut, G. (2019). A new unifying account of the roles of neuronal entrainment. *Current Biology*, *29*(18), R890–R905.

Lakatos, Peter, Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, *94*(3), 1904–1911.

Lehongre, K., Morillon, B., Giraud, A.-L., & Ramus, F. (2013). Impaired auditory sampling in dyslexia: further evidence from combined fMRI and EEG. *Frontiers in Human Neuroscience*, *7*, 454.

LeMay, M., & Culebras, A. (1972). Human brain--morphologic differences in the hemispheres demonstrable by carotid arteriography. *The New England Journal of Medicine*, *287*(4), 168–170.

Lerner, Y., Honey, C. J., Katkov, M., & Hasson, U. (2014). Temporal scaling of neural responses to compressed and dilated natural speech. *Journal of Neurophysiology*, *111*(12), 2433–2444.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, *31*(8), 2906–2915.

Levine, R. A., & McGaffigan, P. M. (1983). Right-left asymmetries in the human brain stem: Auditory evoked potentials. *Electroencephalography and Clinical Neurophysiology*, *55*(5), 532–537.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing …*.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Liang, L., Lu, T., & Wang, X. (2002). Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. *Journal of Neurophysiology*, *87*(5), 2237–2261.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of

speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358–368.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Liégeois-Chauvel, C., Giraud, K., Badier, J. M., Marquis, P., & Chauvel, P. (2001). Intracerebral evoked potentials in pitch perception reveal a functional asymmetry of the human auditory cortex. *Annals of the New York Academy of Sciences*, *930*, 117–132.

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

Liégeois-Chauvel, C., Musolino, A., Badier, J. M., Marquis, P., & Chauvel, P. (1994). Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. *Electroencephalography and Clinical Neurophysiology*, *92*(3), 204–214.

Liegeois-Chauvel, C., Musolino, A., & Chauvel, P. (1991). Localization of the primary auditory area in man. *Brain: A Journal of Neurology*, *114 (Pt 1A)*, 139–151.

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

Lubinus, C., Orpella, J., Keitel, A., Gudi-Mindermann, H., Engel, A. K., Roeder, B., & Rimmele, J. M. (2021). Data-Driven Classification of Spectral Profiles Reveals Brain Region-Specific Plasticity in Blindness. *Cerebral Cortex*, *31*(5), 2505–2522.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010.

Lyon, R. F., & Katsiamis, A. G. (2010). History and future of auditory filter models.

Proceedings of 2010 IEEE ….

Malone, B. J., Scott, B. H., & Semple, M. N. (2010). Temporal codes for amplitude contrast in auditory cortex. *The Journal of Neuroscience*, *30*(2), 767–784.

Marchesotti, S., Nicolle, J., Merlet, I., Arnal, L. H., Donoghue, J. P., & Giraud, A.-L. (2020). Selective enhancement of low-gamma activity by tACS improves phonemic processing and reading accuracy in dyslexia. *PLoS Biology*, *18*(9), e3000833.

Marie, D., Maingault, S., Crivello, F., Mazoyer, B., & Tzourio-Mazoyer, N. (2016). Surface-Based Morphometry of Cortical Thickness and Surface Area Associated with Heschl's Gyri Duplications in 430 Healthy Volunteers. *Frontiers in Human Neuroscience*, *10*, 69.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review*, *101*(4), 653–675.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Marti, Sébastien, & Dehaene, S. (2017). Discrete and continuous mechanisms of temporal selection in rapid visual streams. *Nature Communications*, *8*(1), 1955.

Martin, A. E. (2016). Language processing as cue integration: grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, *7*, 120.

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427.

Marti, S, King, J. R., & Dehaene, S. (2015). Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. *Neuron*, *88*(6), 1297–1307.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., et al. (2020). Camembert: a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7203–7219).

Presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics.

Massaro, D. W. (1974). Perceptual units in speech recognition. *Journal of experimental psychology*, *102*(2), 199–208.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McGettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: what's wrong, what's right and what's left? *Trends in Cognitive Sciences*, *16*(5), 269–276.

McGrew, W. C., & Marchant, L. F. (1999). Laterality of hand use pays off in foraging success for wild chimpanzees. *Primates; Journal of Primatology*, *40*(3), 509–513.

Medland, S. E., Duffy, D. L., Wright, M. J., Geffen, G. M., Hay, D. A., Levy, F., van-Beijsterveldt, C. E. M., et al. (2009). Genetic influences on handedness: data from 25,732 Australian and Dutch twin families. *Neuropsychologia*, *47*(2), 330–337.

Mehler, J., & Hayes, R. W. (1981). The role of syllables in speech processing: infant and adult data [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *295*(1077), 333–352.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, *58*(4), 880–883.

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

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *The European Journal of Neuroscience*, *48*(7), 2609–2621.

Meyer, M., Liem, F., Hirsiger, S., Jäncke, L., & Hänggi, J. (2014). Cortical surface area and

cortical thickness demonstrate differential structural asymmetry in auditory-related areas of the human cortex. *Cerebral Cortex*, *24*(10), 2541–2552.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the …*.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. *Phonetica*, *41*(4), 215–225.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Mišic, B., Betzel, R. F., Griffa, A., de Reus, M. A., He, Y., Zuo, X. N., van den Heuvel, M. P., et al. (2018). Network-Based Asymmetry of the Human Auditory System. *Cerebral Cortex*, *28*(7), 2655–2664.

Moore, Brian C. (2012). An introduction to the psychology of hearing.

Moore, B C. (2008). Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 947–963.

Moore, J. K. (1987). The human auditory brain stem: a comparative view. *Hearing Research*, *29*(1), 1–32.

Moreau, T., & Gramfort, A. (2020). Dicodile: Distributed convolutional dictionary learning. *IEEE Transactions on Pattern Analysis …*.

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neuroscience and Biobehavioral Reviews*, *107*, 136–142.

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(42), E8913–E8921.

Morillon, B., Hackett, T. A., Kajikawa, Y., & Schroeder, C. E. (2015). Predictive motor control of sensory dynamics in auditory active sensing. *Current Opinion in Neurobiology*, *31*, 230–238.

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

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

Morillon, B., Schroeder, C. E., & Wyart, V. (2014). Motor contributions to the temporal precision of auditory attention. *Nature Communications*, *5*, 5255.

Morillon, B., & Schroeder, C. E. (2015). Neuronal oscillations as a mechanistic substrate of auditory temporal prediction. *Annals of the New York Academy of Sciences*, *1337*, 26–31.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *Neuroimage*, *13*(4), 684–701.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, *9*(5–6), 453–467.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: a new French lexical database. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, *36*(3), 516–524.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*(3), 191–243.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Nourski, K. V., & Brugge, J. F. (2011). Representation of temporal sound features in the human auditory cortex. *Reviews in the neurosciences*, *22*(2), 187–203.

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., et al. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *The Journal of Neuroscience*, *29*(49), 15564–15574.

Nourski, K. V., Steinschneider, M., Rhone, A. E., Kovach, C. K., Kawasaki, H., & Howard, M. A. (2019). Differential responses to spectrally degraded speech within human auditory cortex: An intracranial electrophysiology study. *Hearing Research*, *371*, 53–65.

Obleser, J. (2021). A quiet innovator: Peter Lakatos (1972–2021). *Nature Neuroscience*.

Ocklenburg, S, Friedrich, P., Fraenz, C., Schlüter, C., Beste, C., Güntürkün, O., & Genç, E. (2018). Neurite architecture of the planum temporale predicts neurophysiological processing of auditory speech. *Science Advances*, *4*(7), eaar6830.

Ocklenburg, Sebastian, & Güntürkün, O. (2012). Hemispheric asymmetries: the comparative view. *Frontiers in Psychology*, *3*, 5.

Ocklenburg, Sebastian, & Güntürkün, O. (2017). The lateralized brain: The neuroscience and evolution of hemispheric asymmetries.

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of phonetics*, *53*, 153–176.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1718–1725.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911.

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology*, *25*(12), 1649–1653.

Partee, B. (1975). Montague grammar and transformational grammar. *Linguistic inquiry*.

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 358–377.

Pastore, M. (2018). Overlapping: a R package for Estimating Overlapping in Empirical Distributions. *The Journal of Open Source Software*, *3*(32), 1023.

Patterson, R. D., & Nimmo-Smith, I. (1987). An efficient auditory filterbank based on the gammatone function. *a meeting of the IOC …*.

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*, 320.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387.

Peelle, J. E., McMillan, C., Moore, P., Grossman, M., & Wingfield, A. (2004). Dissociable patterns of brain activity during comprehension of rapid and syntactically complex speech: evidence from fMRI. *Brain and Language*, *91*(3), 315–325.

Pefkou, M., Arnal, L. H., Fontolan, L., & Giraud, A.-L. (2017). θ-Band and β-Band Neural Activity Reflects Independent Syllable Tracking and Comprehension of

Time-Compressed Speech. *The Journal of Neuroscience*, *37*(33), 7930–7938.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A CROSS-LANGUAGE PERSPECTIVE ON SPEECH INFORMATION RATE. *Language*, *87*(3), 539–558.

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

Penhune, V. B., Zatorre, R., MacDonald, J. D., & Evans, A. C. (1996). Interhemispheric anatomical differences in human primary auditory cortex: probabilistic mapping and volume measurement from magnetic resonance scans. *Cerebral Cortex*, *6*(5), 661–672.

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology*, *1*, 227.

Pesnot Lerousseau, J., Trebuchon, A., Morillon, B., & Schon, D. (2019). Persistent neural entrainment in the human cortex is frequency selective. *BioRxiv*.

Pesnot Lerousseau, J., Trébuchon, A., Morillon, B., & Schön, D. (2021). Frequency selectivity of persistent cortical oscillatory responses to auditory rhythmic stimulation. *The Journal of Neuroscience*, *41*(38), 7991–8006.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(9), 3526–3529.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

Pirilä, T., Jounio-Ervasti, K., & Surri, M. (1992). Left-right asymmetries in hearing threshold levels in three age groups of a random population. *Audiology*.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. *The Journal of the Acoustical Society of America*, *61*(5), 1352–1361.

Poeppel, David, & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews. Neuroscience*, *21*(6), 322–334.

Poeppel, David, & Hickok, G. (2004). Towards a new functional anatomy of language. *Cognition*, *92*(1–2), 1–12.

Poeppel, David, Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1071–1086.

Poeppel, David, Overath, T., Popper, A. N., & Fay, R. R. (Eds.). (2012). *The human auditory cortex*. Springer handbook of auditory research (Vol. 43). New York, NY: Springer New York.

Poeppel, David. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' *Speech communication*, *41*(1), 245–255.

Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive science*, *25*(5), 679–693.

Prete, G., Marzoli, D., Brancucci, A., & Tommasi, L. (2016). Hearing it right: Evidence of hemispheric lateralization in auditory imagery. *Hearing Research*, *332*, 80–86.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425.

Quinn, A. J., Lopes-Dos-Santos, V., Huang, N., Liang, W.-K., Juan, C.-H., Yeh, J.-R., Nobre, A. C., et al. (2021). Within-cycle instantaneous frequency profiles report oscillatory waveform dynamics. *Journal of Neurophysiology*, *126*(4), 1190–1208.

Racine, J. S. (2012). RStudio: A Platform-Independent IDE for R and Sweave. *Journal of Applied Econometrics*, *27*(1), 167–172.

Raichle, M. E., & Gusnard, D. A. (2002). Appraising the brain's energy budget. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(16), 10237–10239.

Rasmussen, T., & Milner, B. (1977). The role of early left-brain injury in determining lateralization of cerebral speech functions. *Annals of the New York Academy of Sciences*, *299*, 355–369.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724.

Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiology & Neuro-Otology*, *3*(2–3), 86–103.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? . *Journal of Experimental Psychology. Human Perception and Performance*, *18*(3), 849–860.

Reed, C. M., & Durlach, N. I. (1998). Note on information transfer rates in human communication. *Presence: Teleoperators and Virtual Environments*, *7*(5), 509–518.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 978–996.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–949.

Riad, R., Karadayi, J., Bachoud-Lévi, A.-C., & Dupoux, E. (2021). Learning spectro-temporal representations of complex sounds with parameterized neural networks. *The Journal of the Acoustical Society of America*, *150*(1), 353.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*,

22(11), 1761–1770.

Ringo, J. L., Doty, R. W., Demeter, S., & Simard, P. Y. (1994). Time is of the essence: a conjecture that hemispheric specialization arises from interhemispheric conduction delay. *Cerebral Cortex*, *4*(4), 331–343.

Rivier, F., & Clarke, S. (1997). Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. *Neuroimage*, *6*(4), 288–304.

Rogalsky, C., & Hickok, G. (2011). The role of Broca's area in sentence comprehension. *Journal of Cognitive Neuroscience*, *23*(7), 1664–1680.

Rogers, L. J., Zucca, P., & Vallortigara, G. (2004). Advantages of having a lateralized brain. *Proceedings. Biological Sciences / the Royal Society*, *271 Suppl 6*, S420-2.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *336*(1278), 367–373.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494.

R core, T. (2020). *R: A Language and Environment for Statistical Computing*. Computer software, Vienna, Austria: R Foundation for Statistical Computing.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, *398*(6730), 760.

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

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105.

Saleh, M., Reimer, J., Penn, R., Ojakangas, C. L., & Hatsopoulos, N. G. (2010). Fast and slow oscillations in human primary motor cortex predict oncoming behaviorally relevant cues. *Neuron*, *65*(4), 461–471.

Sanabria-Diaz, G., Melie-García, L., Iturria-Medina, Y., Alemán-Gómez, Y., Hernández-González, G., Valdés-Urrutia, L., Galán, L., et al. (2010). Surface area and cortical thickness descriptors reveal different attributes of the structural human brain networks. *Neuroimage*, *50*(4), 1497–1510.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Computational Biology*, *10*(1), e1003412.

Saoud, H., Josse, G., Bertasi, E., Truy, E., Chait, M., & Giraud, A.-L. (2012). Brain-speech alignment enhances auditory cortical responses and speech perception. *The Journal of Neuroscience*, *32*(1), 275–281.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103–113.

Schmidt, F., Chen, Y.-P., Keitel, A., Roesch, S., Hannemann, R., Serman, M., Hauswald, A., et al. (2021). Neural speech tracking shifts from the syllabic to the modulation rate of speech as intelligibility decreases. *BioRxiv*.

Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J.-L., & Besson, M. (2010). Similar cerebral networks in language, music and song perception. *Neuroimage*, *51*(1), 450–461.

Schönwiesner, M., & Zatorre, R. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(34), 14611–14616.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N. G., Tenenbaum, J. B., et al. (2020). Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, *12*(3), 106–113.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*(1), 9–18.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of neurolinguistics*, *25*(5), 336–354.

Scott, S. K., & McGettigan, C. (2013). Do temporal processes underlie left hemisphere dominance in speech perception? *Brain and Language*, *127*(1), 36–45.

Seldon, H. L. (1981). Structure of human auditory cortex. I. Cytoarchitectonics and dendritic distributions. *Brain Research*, *229*(2), 277–294.

Semple, S., Ferrer-i-Cancho, R., & Gustison, M. L. (2021). Linguistic laws in biology. *Trends in Ecology & Evolution*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Sigman, M., & Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. *The Journal of Neuroscience*, *28*(30), 7585–7598.

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of*

America, *114*(6 Pt 1), 3394–3411.

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language*, *164*, 77–105.

Skipper, J. I., & Hasson, U. (2017). A core speech circuit between primary motor, somatosensory, and auditory cortex: Evidence from connectivity and genetic descriptions. *BioRxiv*.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387–2399.

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*(7079), 978–982.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience*, *32*(25), 8443–8453.

Spitzer, B., & Haegens, S. (2017). Beyond the status quo: A role for beta oscillations in endogenous content (re)activation. *eNeuro*, *4*(4).

Stanovich, K. E., & West, R. F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory & Cognition*, *7*(2), 77–85.

Steinmann, S., Leicht, G., Andreou, C., Polomac, N., & Mulert, C. (2017). Auditory verbal hallucinations related to altered long-range synchrony of gamma-band oscillations. *Scientific Reports*, *7*(1), 8401.

Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America*, *55*(3), 653–659.

Stevens, K. N. (2000). *Acoustic Phonetics*. The MIT Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872–1891.

Stilp, C. E., Kiefte, M., Alexander, J. M., & Kluender, K. R. (2010). Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. *The Journal of the Acoustical Society of America*, *128*(4), 2112–2126.

Strauß, A., & Schwartz, J.-L. (2017). The syllable in the light of motor skills and neural oscillations. *Language, cognition and neuroscience*, *32*(5), 562–569.

Studdert-Kennedy, M. (1986). Some developments in research on language behavior. *Behavioral and social science: Fifty years …*.

Sun, Y., Michalareas, G., & Poeppel, D. (2021). The impact of phase entrainment on auditory detection is highly variable: Revisiting a key finding. *The European Journal of Neuroscience*.

Suppes, P., Lu, Z. L., & Han, B. (1997). Brain wave recognition of words. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14965–14969.

Talairach, J., Bancaud, J., Bonis, A., Szikla, G., & Tournoux, P. (1962). Functional stereotaxic exploration of epilepsy. *Stereotactic and Functional Neurosurgery*, *22*(3–5), 328–331.

Taylor, W. L. (1953). “cloze procedure”: A new tool for measuring readability. *Journalism Bulletin*, *30*(4), 415–433.

Ten Oever, S., & Martin, A. E. (2021). An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions. *eLife*, *10*.

Ten Oever, S., Schroeder, C. E., Poeppel, D., van Atteveldt, N., Mehta, A. D., Mégevand, P., Groppe, D. M., et al. (2017). Low-Frequency Cortical Oscillations Entrain to Subthreshold Rhythmic Auditory Stimuli. *The Journal of Neuroscience*, *37*(19), 4903–4912.

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic

model of language neurobiology. *Brain and Language*, *162*, 60–71.

Turkeltaub, P. E., & Coslett, H. B. (2010). Localization of sublexical speech perception components. *Brain and Language*, *114*(1), 1–15.

Tzourio-Mazoyer, N., & Mazoyer, B. (2017). Variations of planum temporale asymmetries with Heschl's Gyri duplications and association with cognitive abilities: MRI investigation of 428 healthy volunteers. *Brain Structure & Function*, *222*(6), 2711–2726.

Tzourio-Mazoyer, N., Perrone-Bertolotti, M., Jobard, G., Mazoyer, B., & Baciu, M. (2017). Multi-factorial modulation of hemispheric specialization and plasticity for language in healthy and pathological conditions: A review. *Cortex*, *86*, 314–339.

Ueda, K., Nakajima, Y., Ellermeier, W., & Kattner, F. (2017). Intelligibility of locally time-reversed speech: A multilingual comparison. *Scientific Reports*, *7*(1), 1782.

Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., & Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. *The Journal of Neuroscience*, *32*(26), 9089–9102.

Vallortigara, G. (2006). The evolutionary psychology of left and right: costs and benefits of lateralization. *Developmental Psychobiology*, *48*(6), 418–427.

Van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *Journal of Philosophy*, *92*(7), 345–381.

Varela, F., Lachaux, J. P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nature Reviews. Neuroscience*, *2*(4), 229–239.

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976.

Versfeld, N. J., & Dreschler, W. A. (2002). The relationship between the intelligibility of

time-compressed speech and speech in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, *111*(1 Pt 1), 401–408.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Voss, R. P. (1975). I/F NOISE"" IN MUSIC and SPEECH.

Vosskuhl, J., Strüber, D., & Herrmann, C. S. (2018). Non-invasive Brain Stimulation: A Paradigm Shift in Understanding Brain Oscillations. *Frontiers in Human Neuroscience*, *12*, 211.

Wada, J. A. (1969). Interhemispheric sharing and shift of cerebral speech function. *Excerpta Medica International Congress Series*, 296.

Wallace, M. N., Johnston, P. W., & Palmer, A. R. (2002). Histochemical identification of cortical areas in the auditory region of the human brain. *Experimental Brain Research*, *143*(4), 499–508.

Wang, X. (2007). Neural coding strategies in auditory cortex. *Hearing Research*, *229*(1–2), 81–93.

Warren, R. M., & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception & Psychophysics*, *9*(3), 358–362.

Warrier, C., Wong, P., Penhune, V., Zatorre, R., Parrish, T., Abrams, D., & Kraus, N. (2009). Relating structure to function: Heschl's gyrus and acoustic processing. *The Journal of Neuroscience*, *29*(1), 61–69.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley interdisciplinary reviews. Cognitive science*, *3*(3), 387–401.

Whittington, J. C. R., & Bogacz, R. (2019). Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, *23*(3), 235–250.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Presented at the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Stroudsburg, PA, USA: Association for Computational Linguistics.

Xiang, J., Poeppel, D., & Simon, J. Z. (2013). Physiological evidence for auditory modulation filterbanks: cortical responses to concurrent modulations. *The Journal of the Acoustical Society of America*, *133*(1), EL7-12.

Yin, P., Johnson, J. S., O'Connor, K. N., & Sutter, M. L. (2011). Coding of amplitude modulation in primary auditory cortex. *Journal of Neurophysiology*, *105*(2), 582–600.

Zaehle, T., Lenz, D., Ohl, F. W., & Herrmann, C. S. (2010). Resonance phenomena in the human auditory cortex: individual resonance frequencies of the cerebral cortex determine electrophysiological responses. *Experimental Brain Research*, *203*(3), 629–635.

Zalta, A., Petkoski, S., & Morillon, B. (2020). Natural rhythms of periodic temporal attention. *Nature Communications*, *11*(1), 1051.

Zatorre, R., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46.

Zatorre, R., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, *11*(10), 946–953.

Zatorre, R. (1989). Perceptual asymmetry on the dichotic fused words test and cerebral speech lateralization determined by the carotid sodium amytal test. *Neuropsychologia*, *27*(10), 1207–1219.

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., et al. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of*

the National Academy of Sciences of the United States of America, *102*(7), 2293–2298.

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The Learning Signal in Perceptual Tuning of Speech: Bottom Up Versus Top-Down Information. *Cognitive science*, *45*(3), e12947.

Zhou, H., Melloni, L., Poeppel, D., & Ding, N. (2016). Interpretations of frequency domain analyses of neural entrainment: periodicity, fundamental frequency, and harmonics. *Frontiers in Human Neuroscience*, *10*, 274.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*, *77*(5), 980–991.

Zipf, G. K. (1935). The psycho-biology of language: An introd. to dynamic philology. *The psycho-biology of language: An introd. to dynamic philology*.

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Current Biology*, *28*(3), 401-408.e5.

Zou, J., Xu, C., Luo, C., Jin, P., Gao, J., Li, J., Gao, J., et al. (2021). θ-Band Cortical Tracking of the Speech Envelope Shows the Linear Phase Property. *eNeuro*, *8*(4).

Zündorf, I. C., Lewald, J., & Karnath, H.-O. (2016). Testing the dual-pathway model for auditory processing in human cortex. *Neuroimage*, *124*(Pt A), 672–681.