Voice mixology at a cocktail party : Combining behavioural and neural tracking for speech segregation

Moïra-Phoebé Huet

N°d'ordre NNT : 2020LYSEI070

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
Institut National des Sciences Appliquées

Ecole Doctorale 162
Mécanique, Energétique, Génie civil, Acoustique

Spécialité de doctorat : Acoustique

Soutenue publiquement le 17/09/2020, par :
Moïra-Phoebé Huet

Voice mixology at a cocktail party

Combining behavioural and neural tracking for speech segregation

---

Devant le jury composé de :

Pressnitzer Daniel	Directeur de Recherche, ENS Paris	Président
McGettigan Carolyn	Prof., University College London	Rapporteuse
Meunier Fanny	Directrice de Recherche, Université Côté d'Azur	Rapporteuse
Parizet Etienne	Prof., INSA Lyon	Directeur de thèse
Gaudrain Etienne	Chargé de Recherche, CRNL Lyon	Co-encadrant

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr** INSA : R. GOURDON	**M. Stéphane DANIELE** Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec. : M.C. HAVGOUDOUKIAN **ecole-doctorale.eea@ec-lyon.fr**	**M. Gérard SCORLETTI** École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 **gerard.scorletti@ec-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES **secretariat.e2m2@univ-lyon1.fr**	**M. Philippe NORMAND** UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://www.ediss-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE **secretariat.ediss@univ-lyon1.fr**	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Curien - 3ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tel : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 **infomaths@univ-lyon1.fr**	**M. Hamamache KHEDDOUCI** Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tel : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **ed.materiaux@insa-lyon.fr**	**M. Jean-Yves BUFFIÈRE** INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 **jean-yves.buffiere@insa-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **mega@insa-lyon.fr**	**M. Jocelyn BONJOUR** INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **http://ed483.univ-lyon2.fr** Sec. : Véronique GUICHARD INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 **veronique.cervantes@univ-lyon2.fr**	**M. Christian MONTES** Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Voice mixology at a cocktail party

Combining behavioural and neural tracking for speech segregation

Moïra-Phoebé Huet

Supervisors: Prof. Etienne Parizet
Dr. Etienne Gaudrain

This dissertation is submitted for the degree of
Doctor of Acoustics

December 2020

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

UNIVERSITÉ DE LYON

LABEX CELYA UNIVERSITÉ DE LYON

Supervisors
Prof. Etienne Parizet
Dr. Etienne Gaudrain

Jury members
Carolyn McGettigan (rapportor)
Fanny Meunier (rapportor)
Daniel Pressnitzer (president)

The work presented in this thesis was performed at the Cognition Auditive et Psychoacoustic team (CAP), Center of Research in Neurosciences of Lyon (CRLN), Université Claude-Bernard Lyon 1 and at the Laboratory of Vibration and Acoustics (LVA), National Institute of Applied Sciences (INSA Lyon).

The studies in this thesis were funded by the LABEX CeLyA (Centre Lyonnais d'Acoustique) (ANR-10-LABX-0060) of Université de Lyon, within the program « Investissements d'Avenir » (ANR-16-IDEX-0005).

Acknowledgements - Remerciements

Je tiens à remercier chaleureusement mes deux directeurs de thèse, Etienne Parizet et Etienne Gaudrain, avec qui j'ai pris beaucoup de plaisir à travailler. Merci pour leur confiance, leur patience et leurs encouragements. Merci pour nos longs échanges qui ont été formateurs sur un large éventail de sujets et ont permis l'aboutissement de ce travail. À jamais merci aussi de m'avoir fait découvrir le monde fantastique de la psychoacoustique. Je remercie également vivement Christophe Micheyl pour ses conseils avisés, ses enseignements et son importante contribution à ce travail.

I would like to thank all the committee members for agreeing to participate in the evaluation of this thesis: Carolyn McGettigan and Fanny Meunier (who both agreed to review this manuscript) and Daniel Pressnitzer.

I could never have finished this project without the support of my family, especially my parents, and the companionship of many people who became my friends during this doctoral journey. Special thanks to Anna Fiveash (for all our conversations over coffee and with whom I enjoyed collaborating), Lizette Heine (who provided insightful advice and who was always there for the good and the bad times), Jackson Graves (for his precious help, his passion for languages and his dear friendship) and Sofia Gambaro (che mi ha dato un feedback chimico). Enfin, un immense merci à Léo Papet, mon frère de thèse, avec qui j'ai adoré franchir chaque étape de cette folle aventure, pour son inébranlable bonne humeur, son hospitalité et tous ses encouragements.

Je voudrais également remercier les équipes CAP et LVA pour leur accueil. Merci à Nicolas Grimault, Fabien Perrin, Barbara Tillmann, Yohanna Lévêque et Marie Avillac pour leurs nombreux conseils. Un merci particulier à Alexandra Corneyllie pour son inestimable aide dans la mise en place des expériences, et pour m'avoir fait rêver tout au long de ces années. Un grand merci à Julie Thevenet, Margot Bouhon et Samar Dimachki qui ont contribué à rendre notre bureau

agréable et convivial. Merci aussi à Agathe Pralus, Lucile Rey, Xiaoxia Sun pour leur gentillesse et leur soutien. Merci à Samuel Garcia de m'avoir présenté Python et Linux, ainsi qu'à Hervé Hugueney pour son soutien technique. Merci à Brigitte Teissier, Florence Tamissa-Léger et Romain Saroli pour leur soutien logistique. Merci aussi à Nathalie Loriot pour sa redoutable efficacité et son support administratif.

Puis, je voudrais remercier le LabEx CeLyA d'avoir financé cette thèse. Merci à Carine Zambardi et Agnès Delebassee-Nabet pour leur amabilité et leur aide administrative. Merci également au département Fedora d'avoir toujours su répondre à toutes mes questions. Merci aussi à l'espace Ulys pour m'avoir guidée à travers le labyrinthe du système français.

Enfin surtout, le plus grand des mercis à Clémence pour son soutien quotidien, ses valeureux services de cobaye et nos nombreuses chansons qui auront rythmé, et enjolivé, cette odyssée lyonnaise.

Abstract

It is not always easy to follow a conversation in a noisy environment. In order to discriminate two speakers, we have to mobilize many perceptual and cognitive processes to maintain attention on a target voice and avoid shifting attention to the background. In this dissertation, the processes underlying speech segregation are explored through behavioural and neurophysiological experiments.

In a preliminary phase, the development of an intelligibility task – the Long-SWoRD test – is introduced. This protocol allows participants to benefit from cognitive resources, such as linguistic knowledge, to separate two talkers in a realistic listening environment. The similarity between the two speakers, and thus by extension the difficulty of the task, was controlled by manipulating the acoustic parameters of the target and masker voices.

In a second phase, the performance of the participants on this task is evaluated through three behavioural and neurophysiological studies (EEG). Behavioural results are consistent with the literature and show that the distance between voices, spatialisation cues, and semantic information influence participants' performance. Neurophysiological results, analysed with temporal response functions (TRF), indicate that the neural representations of the two speakers differ according to the difficulty of listening conditions. In addition, these representations are constructed more quickly when the voices are easily distinguishable.

It is often presumed in the literature that participants' attention remains constantly on the same voice. The experimental protocol presented in this work provides the opportunity to retrospectively infer when participants were listening to each voice. Therefore, in a third stage, a combined analysis of this attentional information and EEG signals is presented. Results show that information about attentional focus can be used to improve the neural representation of the attended voice in situations where the voices are similar.

Table of contents

List of Figures

List of Tables

Nomenclature

Greek Symbols

Φ	Semantic context measure

Other Symbols

F0	Fundamental frequency
R	Stimulus-reconstruction evaluation *or* Neural tracking accuracy

Acronyms / Abbreviations

AAD	Auditory attention decoding
ASA	Auditory Scene Analysis
CRM	Coordinate Response Measure *(test)*
DNN	Deep Neural Networks
EEG	Electroencephalography
ERP	Event-Related Potential
fMRI	functional Magnetic Resonance Imagining
GAM	Generalized Additive Models
gLMM	generalized Linear Mixed Model
ISE	Irrelevant Sound Effect
LMM	Linear Mixed Model
Long-SWoRD	Long Selective Word Recognition Discrimination *(test)*

MEG	Magnetoencephalography
STRF	Spectro-Temporal Receptive Fields *and also* Spectro-Temporal Response Function
TMR	Target-to-Masker Ratio
TRF	Temporal Response Function
VTL	Vocal Tract Length

Chapter 1

Cocktail Party Phenomenon: theoretical, behavioural and neurophysiological insights

1.1 Introduction

At the vernissage of an art exhibition, some people are talking about the paintings. The room is full of sounds: voices, fragments of conversations, jazzy background music and clinking glasses; here someone is ecstatic about the beauty of a piece of art; there, another tells a joke... And yet, the young woman who is visiting the exhibition while chatting with her friends does not seem to be bothered by the background noise as if it did not exist. Nothing extraordinary... and yet... Without this young woman even noticing it, her brain performs a masterful trick: filtering her friends' voices off from all the noises and conversations.

The ability of humans to listen in extreme acoustic conditions has been studied for more than sixty years and we still do not know exactly how the normal auditory system parses these complex scenes. Answering this question is crucial as it is a situation which can lead to real-life problems such as poor communication over unreliable communication lines. When research on this topic began, complex acoustic scenes were noisy telephone or radio-telecommunication; today, complex acoustic scenes include an unstable internet connection during a Skype meeting.

Challenges encountered in such situations can be accentuated for people with hearing loss or cochlear implants who may experience difficulties focusing on conversation in noisy environments.

Colin Cherry (1953), a British cognition specialist, was one of the first to be interested in what is called nowadays the *"Cocktail Party Problem"* (or "Cocktail Party Effect"). Cherry's work acted as a starting point for different lines of research. First, it has had a major influence on *selective attention* research with: (1) influential early filter models (*e.g.* Broadbent, 1958); (2) understanding how auditory attention (and more broadly memory and cognition) affects speech perception (*e.g.* Cowan & Wood, 1997) and, more recently, in (3) the identification of the neural correlates of the auditory selective attention (*e.g.* Mesgarani & Chang, 2012). Another major line of research has addressed the question of the *separation of two speech streams* (for a review, see Bregman, 1994). This latter research axis has been carried by Albert Bregman who has introduced in his famous book a novel term for people's efforts to solve these cocktail party situations: the *auditory scene analysis (ASA)*.

A common case of the cocktail party effect is the simpler situation where only two speakers are competing (also referred as *speech-on-speech*). Although this represents a simplified version of the real life cocktail party effect, tremendous work has been done to deeper understand how humans can easily attend to one speaker and ignore the other one. In those studies, researchers are interested in the ability of listeners to select *a target speech* stream while ignoring another voice (referred to as *the masker speech stream*).

In the following sections, the key elements involved in a the speech-on-speech situation will be described. First, the contributions of psychoacoustics and, more broadly, of auditory scene analysis will be discussed in Section 1.2, as this is one of the major research areas that preceded Cherry's work. Then, Section 1.3 will focus on research in the field of attention and more particularly on the latest advances in neural tracking. Finally, the peculiarities of language studies will be highlighted in Section 1.4.

1.2 Formation and selection of auditory objects

The sounds that we hear come generally from different acoustic sources. These sources emit time-varying acoustic waves that propagate to the entrance of the

ear where they combine to form a single sound signal called a *mixture* (Bregman, 1994). Figure 1.1 details the auditory pathways in the brain, from the cochlea (which transform the stimulus into neural signals) to the auditory cortex. However, we usually do not perceive this mixture as a single signal and our brain decomposes the signal into sound objects. These sound object representations correspond to the sounds that ideally match each of the acoustic sources.

The process of estimating which mixture components come from the same source is the first task performed to successfully negotiate the challenge of listening to a talker while ignoring background noise or other talkers. This process is called auditory object *formation* (Shinn-Cunningham, Best, & Lee, 2017). Once the auditory objects are formed from a sound mixture, the listener needs to focus their attention on that source. This second task is called auditory object *selection* (Shinn-Cunningham et al., 2017). Together, these processes set up different aspects of how the cocktail party problems can be solved. The next Section 1.2.1 will detail the auditory object formation, while Section 1.2.2 will describe the auditory object selection.

1.2.1 Segregation of two streams

According to Bregman (1994), the role of perception is to derive a useful representation of reality from the sensory inputs. To do so, the auditory system has to perceptually organize the auditory objects that come together and those that do not. These processes of *integration* and *segregation* contribute to create coherent events of sounds.

Auditory objects have several characteristics and general features. Thus, auditory objects have spectrotemporal properties that make them separable from other auditory objects. According to Bregman (1994), the principle of similarity is essential for grouping sounds together. For instance, two sounds that share similar acoustic features are more likely to come from the same source.

This grouping process operates at two levels (Shinn-Cunningham et al., 2017). On the one hand, auditory objects can be grouped when they share similar sound energy. On the other hand, these auditory objects can also be grouped across longer time scale, forming what Bregman referred to as “streams”.

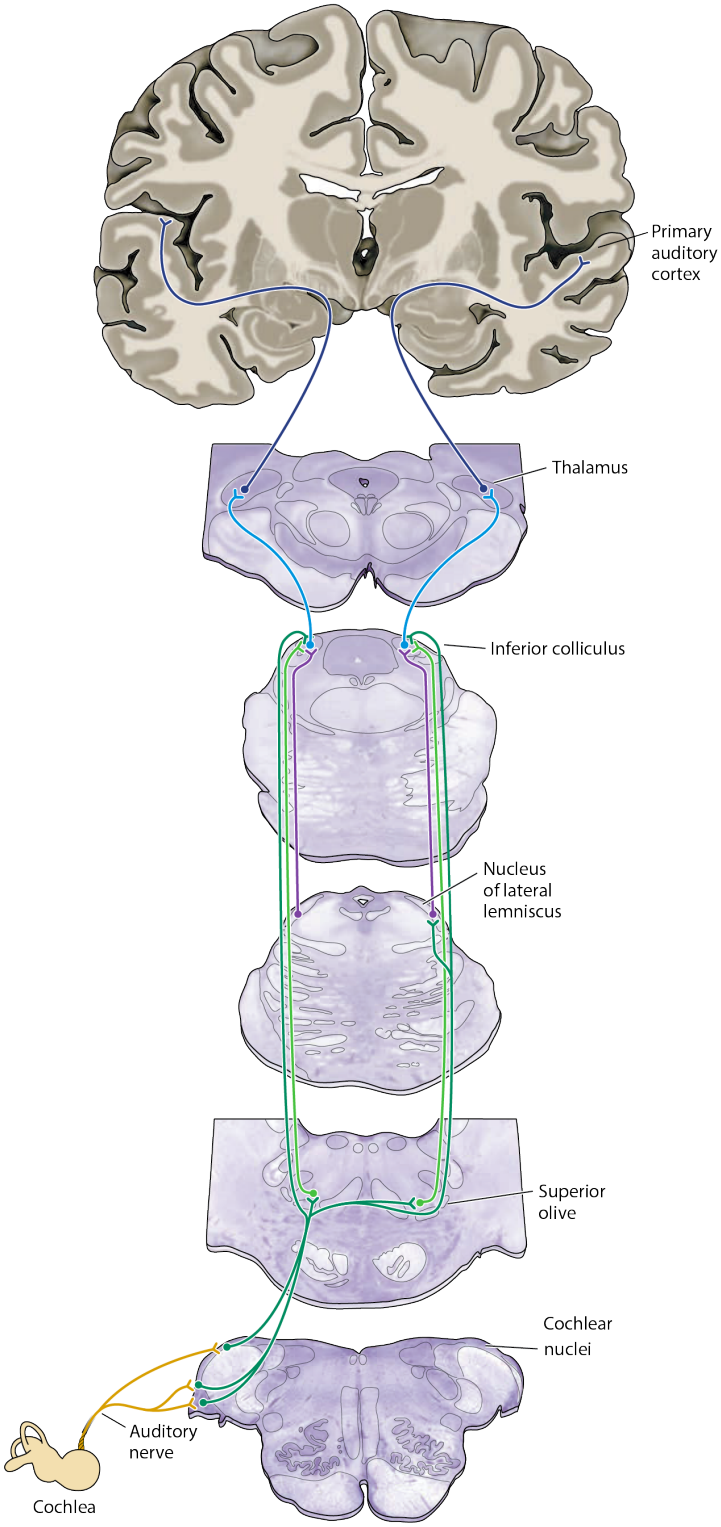

Figure 1.1 The auditory pathways (modified from Purves, 2018).

Simultaneous segregation

According to the spectrotemporal proximity rule, sounds tend to be grouped if they are close together and continuous in time and/or in frequency (Shinn-Cunningham et al., 2017). For instance, this grouping process rests on auditory cues such as pitch, harmonicity, common onset/offset and spatial location.

In speech, simultaneous segregation has been mainly studied through the "double vowel" paradigm. In those studies, two synthetic vowels are presented to listeners who have the task of identifying the vowels. Difference in pitch and common onset/offset are the main features studied (for reviews, see de Cheveigné, 1999; Micheyl & Oxenham, 2010). Moreover, Micheyl and Oxenham (2010) report that neural responses to concurrent vowels have been recorded at various levels of the auditory system. For instance, measures of the electrophysiological responses of single neurons (also called *single-unit studies*) have identified potential cues for the pitch separation of simultaneous vowels at the level of the auditory nerve and the ventral cochlear nucleus. *Electroencephalography* (EEG)[1] and *functional magnetic resonance imaging* (fMRI)[2] studies (Alain, Reinke, He, Wang, & Lobaugh, 2005; Alain, Reinke, McDonald, et al., 2005) provided information concerning the brain activity at higher levels. Listening to concurrent vowels reveals neural activation in thalamus and auditory cortex. These studies suggested that segregation of concurrent vowels in the different levels of the auditory hierarchy is automatic or effortless.

Most of the simultaneous segregation studies use rather simple auditory stimuli such as syllables. However, they do not necessarily reflect what happens when listening to a cocktail party mixture. The fact that individual syllables can be perceived to does not always compensate for the real challenge of tracking the stream of such syllables over time (Shinn-Cunningham et al., 2017).

Sequential segregation

Two sounds can also be assigned to different sources if the segregation occurs across a longer time scale. The auditory system must organize sound elements scattered in frequency and time into coherent "streams" in order to identify and,

[1] *Electroencephalography* or *EEG* is a non-invasive technique that records the electrical activity from the brain

[2] *functional Magnetic Resonance imaging or fMRI* is a non-invasive technique that records the brain activity by measuring changes in the oxygenation level of blood in brain cells

hence segregate, sound objects. Higher-order perceptual features such as similarity and continuity are key to this phenomenon, also referred to as *streaming*.

The "ABA streaming" method is widely used to examine the sequential segregation (Grimault, Micheyl, Carlyon, Arthaud, & Collet, 2000; Van Noorden, 1975). In this task, listeners report hearing one or two streams in tone sequences ABA–ABA–..., where A and B are two tones at different frequencies. Frequency separation and a difference of speed between the two tone sequences usually determine most stream segregation. When the frequencies of A and B are very close to each other, listeners are likely to report one stream. When this separation is large, listeners usually report hearing two streams. At intermediate frequency separations, the perception of the listeners may often flip from one stream to two streams. This phenomenon is called "*bistability*" (B. C. J. Moore & Gockel, 2012). Figure 1.2 illustrates these different scenarios.

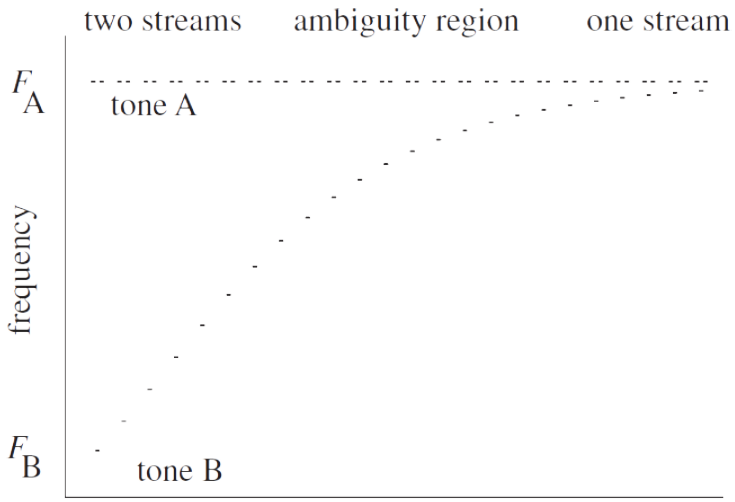

Figure 1.2 The ABA perception is affected by the frequency separation of the tones (modified from B. C. J. Moore & Gockel, 2012).

Studies have looked at the neurophysiological activity involved in the streaming of ABA sequences (for a review, see Bizley & Cohen, 2013). Correlates of sequential streaming can already be observed in the cochlear nucleus (Pressnitzer, Sayles, Micheyl, & Winter, 2008). The main hypothesis is that two sequences of sounds can be streamed when they activate different populations of neurons as a result of a difference in feature of the sound such as the frequency. Moreover, a series of studies on the temporal coherence between different sound features showed that onset synchronicity is also involved in the grouping of stimuli into streams (*e.g.*

Elhilali, Ma, Micheyl, Oxenham, & Shamma, 2009; Micheyl, Kreft, Shamma, & Oxenham, 2013).

Bregman (1978) showed that the tendency to separate two streams increases over the course of a long stimulus. In this construction phase called the *build-up*, the listeners accumulate stochastic evidence that the two streams differ in feature. According to Bregman, it takes a listener about 4 seconds for a sequence of tones to split into streams. This build-up can also emerge over the course of 2 to 10 s, depending on the frequency separation between the two streams (Bregman, 1994). The segregation in streams results from the fact that the stimuli composing a sequence can be categorized on the basis of a feature (or a set of features). It is assumed that the build-up time is the time needed by the auditory system to detect the regularity underlying the structure in which the stimuli will be perceptually organised.

Very little is known about the build-up of more complicated auditory stimuli such as speech. In a recent study, Best, Swaminathan, Kopčo, Roverud, and Shinn-Cunningham (2018) asked participants to listen to digit sequence presented from five locations and showed reductions in confusions between the target and the maskers — hence the build-up — over the course of three to four digits (on the order of two seconds). The authors also hypothesized that the build-up would be more a refinement of selective attention rather than an improvement in segregation.

Just as with simultaneous segregation, many studies of sequential segregation were conducted using short stimuli, such as tone, noise bursts or syllables. However, in the context of cocktail party situations, listeners can rely on a myriad of cues to distinguish a particular talker from competing streams (see Chapter 3 for a review). Along with the physical properties of sounds in the environment, listeners' attention can play a role in streams segregation.

1.2.2 Attention and auditory objects

The role of attention in streams segregation is, still nowadays, a subject of debate. In this section, elements that support the role of attention in the perception of auditory objects will be first presented. Then, evidence that stream segregation can occur in the absence of attention will be exposed through the paradigm of the

"Irrelevant Sound Effect" (B. C. J. Moore & Gockel, 2012). Finally, the question of switching attention will be addressed.

Top-down and bottom-up attention

In the context of an extensive cocktail party situation, it seems unlikely that our cognitive resources analyse what the speaker said as well as all the information coming from the background noise. Instead of such a thorough analysis of the auditory scene, it seems more likely that the selection of one or more speakers is the goal in everyday communication (Shinn-Cunningham et al., 2017).

Listeners can selectively listen to a speaker in an auditory scene by directing their attention. This type of attention is called *top-down attention*. Listeners usually focus on different acoustic features, many of which also influence stream formation. Many studies demonstrated that listeners can focus their attention on the basis of spatial location, pitch, sound level or talker characteristics such as timbre and gender (for a review, see Shinn-Cunningham et al., 2017).

The salience of an auditory stimulus in a mixture can also be enhanced by some factors such as unexpectedness and uniqueness. This type of phenomenon is called *bottom-up attention*. A door that slams suddenly and attracts a listener's attention is a typical example of bottom-up attention. In the context of the cocktail party situation, the sound of our own name that captures our attention, even when uttered by someone we were not listening to, is the most well-known example (Moray, 1959). This latter illustration also reflects the experimental difficulty of studying only the bottom-up salience cues, isolated from the top-down cues: such bottom-up attention, although mainly involuntarily stimulus-driven, is additionally based on the learned importance of that stimulus (Shinn-Cunningham et al., 2017).

The irrelevant sound effect

The "Irrelevant Speech Effect" describes how a speech stream can perturb a short-term visual memory task (Salamé & Baddeley, 1982). Originally called *"Irrelevant Speech Effect"*, this phenomenon has gradually been relabelled the *"Irrelevant Sound Effect"* without having to change its acronym (*ISE*) when it appeared that other signals could also disturb a visual recall task (Ellermeier & Zimmer, 2014). There is a whole panoply of literature that has developed over the last 30 years

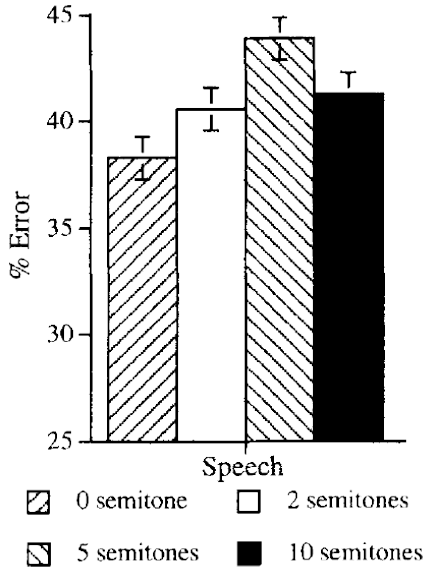

Figure 1.3 Percentage of recall errors in relation to F0 distance between the alternating high and low pitches (modified from Jones et al., 1999).

under the scope of workers' well-being, particularly regarding the suitability of open-space offices (*e.g.* Brocolini, Parizet, & Chevret, 2016; Kostallari, Parizet, Chevret, Amato, & Galy, 2020).

According to B. C. J. Moore and Gockel (2012), the ISE phenomenon can be used to demonstrate that (full) attention is not necessarily required to segregate two streams. Jones, Alford, Bridges, Tremblay, and Macken (1999) used an ABA streaming as an irrelevant sound during a visual serial digit recall task. The sequences of alternating high and low pitch of the ABA task were separated by 0, 2, 5, and 10 semitones. These separations are such that stream segregation should not appear in one condition (0 st) and is expected to happen in at least one condition (10 st). The two remaining conditions (2 and 5 st) should lead to a bistabiliy phenomenon. Figure 1.3 shows the obtained results . When two streams are perceived, listeners make more recall mistakes than when only one stream is perceived, which is consistent with the ISE literature. In the bistability conditions, participants made at least as many mistakes as in the condition where there are two or more perceived streams. Overall, Jones et al. (1999) suggested that streaming processes were at work and that these modulated the disruptive effect of the irrelevant sound. Since participants were not actively attending to the sound, this could suggest that stream segregation can occur outside the focus of the listeners' attention.

It does not seem too far-fetched to assume that the ability to ignore an irrelevant speech signal in a serial recall task would be related to the ability to ignore a masking speech signal in one ear while listening to a different speech signal in the other ear (i.e. a *"dichotic"* listening situation; see Chapter 3 for more information). Comparing behavioural and neurophysiological studies of the two phenomena, Beaman, Bridges, and Scott (2007) concluded that unattended speech signal is processed along two neural pathways, leading to two different behavioural effects. In this framework, the irrelevant speech from an ISE task is largely processed by the right hemisphere whereas, in a dichotic listening task, the unattended speech signal is largely processed by the left hemisphere. The different behavioural results observed across the two paradigms are further highlighted with the operation span task (OSPAN). Moderating effects of working memory capacity can be observed in dichotic listening situations whereas none are present in the ISE paradigm. Beaman et al. (2007) suggest that the neural pathway that processes unattended speech in dichotic listening enables an attentional control process that can be measured with OSPAN. Thus, in dichotic listening, participants with high working memory spans can more easily focus on the task and, by extension, block out the irrelevant message (Conway, Cowan, & Bunting, 2001). From this perspective, the operation span measured by OSPAN appears to be related to a level of attentional control that prevents preventing switching attention from an attended stream to an unattended stream.

Attentional switching

How easily and quickly selective attention can be switched from one auditory stream to another is a question that has interested researchers for many decades (Shinn-Cunningham et al., 2017). Cherry and Taylor (1954) observed a deficit in word repetition when stimuli were presented alternately to the two ears. The time required to disengage and re-engage attention is the cost of switching attention. Another explanation carried by Best, Ozmeral, Kopčo, and Shinn-Cunningham (2008) is that when attention is sustained on one stream, attentional selectivity improves over time. This hypothesis is based on a series of studies (Carlyon, Cusack, Foxton, & Robertson, 2001; Cusack, Decks, Aikman, & Carlyon, 2004) showing that the build-up of stream segregation is reset if attention is diverted away from the target.

In conclusion, several leads seem to emerge on the question of the mandatory involvement of attention in the segregation stream. As explained above, attention from different top-down and bottom-up processes influence streams segregation. However, claiming that segregation of auditory streams (stream formation) and selective attention to a stream (stream selection) are independent phenomena would be wrong. It is much more likely that the two processes are not hierarchically organized but rather occur in parallel, and influence one another and feed back into each other. Finally, some forms of stream segregation can occur in the absence of attention as illustrated by ISE.

The neurophysiology of selective attention, for its part, has seen remarkable advances over the past decade with the emergence of new technologies. Thus, a whole new literature has developed around the study of selective attention, with emphasis in cocktail party situations. The next section focuses on this new field of research called the "neural tracking of speech".

1.3 The neural tracking of speech

In a landmark study, Luo and Poeppel (2007) showed that the phase pattern extracted from brain activity can track and distinguish spoken sentences from each other. The authors hypothesized that the dynamic brain activity is temporally entrained by the dynamic regularities in speech. This study paved the way toward what is nowadays referred to as neural entrainment.

1.3.1 Definitions

The literature on neural entrainment is relatively recent and wide. As a result, it is not always easy to find one's way around without confusing most of the terms. This is why, in a first step, some key notions must be defined and a terminology, agreed upon.

In a narrow sense, motivated from physical principles, entrainment could be defined as the synchronization of rhythms of oscillators (for more details about this definition, see Obleser & Kayser, 2019). Whether the neural process is oscillating in the absence of external stimuli remains unclear and highly debated in studies that refer to "entrainment". A temporal alignment, possibly oscillatory or at least pseudo - rhythmic, of the neural activity and an auditory signal could be

a manifestation of true entrainment in a narrow sense. According to Doelling, Assaneo, Bevilacqua, Pesaran, and Poeppel (2019), what is more likely measured is a cascade of stereotyped, impulse-like evoked responses to a series of sensory inputs. The neural entrainment, in this broader sense, comes down to estimating these impulse responses. To avoid confusion, Obleser and Kayser (2019) recommend using the term *"neural tracking"* to refer to this second definition.

Among tracking measures, regression models and phase coherence are the two most popular methods to link brain activity to stream signals. Regression measures have been widely exploited in cocktail party situations, most notably under the label of *"Auditory Attention Decoding" (AAD)*. These specific measures are therefore presented, in the following section.

1.3.2 Auditory Attention Decoding

From STRF to TRF

Regression models that link songbirds brain activity and naturalistic stimuli, such as animal vocalizations, have been established for single neurons in the early 2000s (Theunissen et al., 2001). A central feature of such models are the *"spectro-temporal receptive fields (STRFs)"*. These STRFs represent the mathematical descriptions of the selectivity of neurons in response to sound events and can be thought of as filters or, decoders, in those models (see Chapter 5 for a more technical description).

Mesgarani, David, Fritz, and Shamma (2009) transposed this method by studying the neural activity of the primary auditory cortex of ferrets with *electrocorticographic* (EcoG)[3] recordings. The working method based on a linear regression is schematically reported in Figure 1.4. Exploiting this approach, the authors managed to reconstruct, at least partially, the speech signal from the EcoG recording. In other words, it became possible to partially reconstruct the sound we hear from the recorded brain activity. A few years later, Pasley et al. (2012) replicated these results in humans and were able to reconstruct a speech stream heard by epileptic listeners equipped with ECoG electrodes.

During the same year, 2012, three different studies implemented the above-mentioned technique in cocktail party situations. Mesgarani and Chang (2012)

[3] *Electrocorticography* or *EcoG* is an invasive technique, usually requiring surgery to implant the electrode grid, which records the electrical brain activity from the brain.

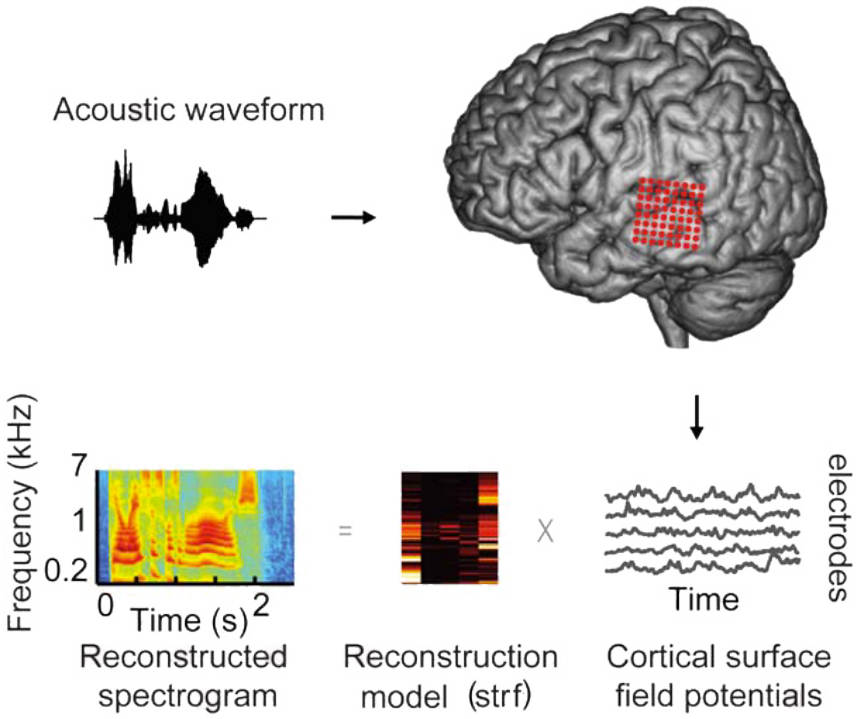

Figure 1.4 Stimulus-reconstruction Method. Participants listen to an acoustic stimulus while their neural activity is recorded (here with an EcoG). Then, a decoder is applied to reconstruct the original sound. The result is a reconstructed sound feature (here a spectrogram) (modified from Pasley et al., 2012).

showed that speech reconstruction, based on cortical responses to the monaural mixture of two speakers, reveal salient sound features of the target speaker: the cortical representation of the target speaker is enhanced relative to that of the masker speaker, almost as if subjects were listening to the target speaker alone. The authors also introduced with this study a classifier that allows to identify the speaker that participants are listening to. This classifier was notably able to recognize the target speaker in 93% of the trials during which the subjects could correctly answer questions about the target speech. This method of speaker identification based on neuronal activity would be labelled, a few years later, *"auditory attention decoding (AAD)"*. However, the cortical locations investigated in this study were constrained by clinical electrode placement and were narrowed to auditory cortex, while it is clear that other brain structures are involved in cocktail party scenarios.

At the same time, Ding and Simon (2012b) applied the linear regression method to *magnetoencephalography* (MEG)[4]. In studies involving non-invasive techniques, the term *spectro-temporal receptive fields* is not directly appropriate. The authors therefore choose to analyse their data with the *spectro-temporal response function*. It is interesting to note that the mathematics is similar as well as the acronym (STRF). Ding and Simon (2012b) estimated the STRF and emphasized the temporal properties of the attentional effects. They found that STRFs have a response peak at the latency of approximately 100 ms that was stronger for the target speaker than for the masker. This latency of 100 ms corresponds to the N1 response which is correlated to the probability of stream segregation (Gutschalk et al., 2005). In a second investigation, the same authors (Ding & Simon, 2012a) showed which brain processes were involved in the separation of two competing speakers and the authors extended their results to reconstruction methods. In accordance with the results of Mesgarani and Chang (2012), they showed that the target stream representation is stronger than the masker representation. Ding and Simon (2012a) also observed a 50 ms post-stimulus latency that did not differentiate between the target and masker streams.

In an EcoG study, Zion Golumbic et al. (2013) drew the link between Mesgarani and Chang (2012), who analysed the high gamma power activity (70-150 Hz), and Ding and Simon (2012a), who the investigated low-frequency neural activity (1-7 Hz),. Zion Golumbic et al.'s results (2013) suggested that attention modulates

[4] *Magnetoencephalography* or *MEG* is a non-invasive technique that records the magnetic field induced by electrical brain activity

the representation of streams in the low-level auditory cortex: the cortical tracking of the target stream is enhanced but the masker stream is still present. On the other hand, the cortical tracking of the masker stream appears not to be detectable in higher-order regions involved in language processing and attentional control regions such as inferior frontal cortex, anterior and inferior temporal cortex, and inferior parietal lobule.

Finally, O'Sullivan et al. (2015) replicated all these results with EEG. Although EEG recordings are much noisier than EcoG or MEG, the authors still managed to obtain an auditory attention decoding accuracy varying in the range of 82-89%. This investigation is a major turning point in the study of selective attention through regression methods since it demonstrates its feasibility with EEG, a much more accessible and less expensive technique than the previous ones. It is from this point on that, to disambiguate the two terms covering "STRF", the term spectro-temporal response function is shortened in the literature to *Temporal Response Function* (TRF).

Challenges and practical applications

The craze for attentional decoding since the publication of O'Sullivan et al. (2015)'s study is at its peak. One of the reasons for this success lies in the potential progress it can generate for the hearing aid technology. If listening and communicating in a babble of multiple talkers might seem relatively easy for normal listeners, it still remains a challenge for hearing-impaired listeners. Hearing aids can facilitate audibility of certain sounds, but do so indiscriminately, regardless of what the user is trying to listen to. If hearing aids could be informed with the target speaker features, they could amplify the target speaker selectively and, hereby, would make it easier for the user to hear them out in crowded environments.

However, there are still many difficulties related to the implementation of this technology. First, the portability of neural recording measurements remains a challenge although it has been greatly improved with the use of EEG. Recent studies (Fiedler et al., 2017; Mirkovic, Bleichner, De Vos, & Debener, 2016) using Concealed EEG Around the Ear, or more simply, a few electrodes placed around the ears, have demonstrated the feasibility of such possible attentional decoding. Second, the attentional decoding remains possible despite artefact complications (Nogueira et al., 2019; Somers, Verschueren, & Francart, 2018; Verschueren, Somers, & Francart, 2019). The next challenge is to reduce the high computational cost

of the auditory attention decoding. In order to limit this computational cost, new calculation methods have emerged, including canonical component analysis (CCA) (de Cheveigné et al., 2018), the use of Markov chains (Geirnaert, Francart, & Bertrand, 2020) and, the exploitation of deep neural networks (DNN) (Akbari, Khalighinejad, Herrero, Mehta, & Mesgarani, 2019; Ciccarelli et al., 2019; de Taillez, Kollmeier, & Meyer, 2018; O'Sullivan et al., 2017).

1.3.3 Intelligibility and brain rhythms

In addition to the practical considerations of the attentional decoding contribution, neural entrainment (in the broad sense) studies provide a better understanding of the implication of intelligibility in the perception of speech.

Peelle, Gross, and Davis (2013) examined how speech-related brain activity was affected by manipulations of the speech intelligibility. With MEG recordings, they showed that neural activity phase-locked to unintelligible speech and that phase locking was enhanced when speech stream was intelligible. These results suggested that the speech tracking does not only depend on acoustic characteristics, but is also affected by listeners' ability to extract linguistic information.

Most recently, studies investigated the role of intelligibility in neural entrainment through *transcranial alternating current stimulation* (tACS)[5]. Riecke, Formisano, Sorger, Başkent, and Gaudrain (2018) modified the cerebral activity with electrical stimulation modulated by a speech envelope and showed modulations of the speech intelligibility. The same year, two other studies (Wilsch, Neuling, Obleser, & Herrmann, 2018; Zoefel, Archer-Boyd, & Davis, 2018) reported similar results. In addition, Zoefel et al. (2018) coupled the tACS to fMRI and reported that electrical manipulation modulates brain activity in the superior temporal gyrus. Altogether, these three studies suggest that entrainment does play a causal role in speech comprehension.

Intelligibility is inherent in communication. The next section briefly discusses the particularities of language carried by a speech signal.

[5] *Transcranial alternating current stimulation* or (tACS) is a neurostimulation technique that delivers an electrical state to the brain via electrodes on the head

1.4 Peculiarities of language

In order to understand language, our brain transforms sound units into syllables which are themselves transformed into words to construct a sentence (Chomsky, 2002). From this sentence will emerge a general sense of context, called semantics, which will help us to understand the message that is intended to be communicated. In this section, we will briefly describe how our linguistic knowledge can change our perception of speech stimuli.

It has been accepted for a long time that listeners benefit from (lexical and semantic) context in language, especially in adverse condition listening (G. A. Miller, Heise, & Lichten, 1951). On the other hand, it is more complicated to determine whether contextual information can contribute more than acoustic cues, but it would seem that, generally, lexical information outweighs acoustic cues under cognitive load (Mattys, Davis, Bradlow, & Scott, 2012).

Phonemic restoration (Warren, 1970) can also show how acoustic cues and context information can influence our perception of speech. The phonemic restoration effect is a perceptual phenomenon in which a listener hears, under certain conditions, sounds actually missing from a speech signal. Usually, a better intelligibility of interrupted speech with periodic silent intervals is observed after these silent intervals are filled with noise. An interplay between bottom-up acoustic cues and top-down cognitive mechanisms is usually observed in phoneme restoration even though the top-down mechanics seem to be more salient. For instance, Clarke, Gaudrain, Chatterjee, and Başkent (2014) showed that linguistic context can influence speech perception to the extent of perceptual voice continuity.

In speech-on-speech studies, another literature has also endeavoured to describe this phenomenon: *energetic and informational masking*. In short, energetic masking occurs when there is at least a partial overlap in time and frequency between a target and a masker signal. Informational masking is defined as a subtraction of energetic masking: it is the effect of the masker when its energetic effect has been accounted for. It typically refers to attention, semantic or cognitive load (Mattys et al., 2012). For instance, Dekerle, Boulenger, Hoen, and Meunier (2014) investigated whether informational masking relies on semantic interference with a background composed of one to four voices and a target. The greater the number of voices in the background, the lower the target intelligibility is. The overall results suggested that informational masking can occur at a semantic level if intelligibility of the masker is sufficient.

Informational masking could be akin to the schema-based segregation of Bregman (1994) who mentioned that our linguistic knowledge could influence stream segregation. Gautreau, Hoen, and Meunier (2015) investigated the effects of listeners' knowledge of the masker on a lexical task. When the masker is spoken in a known language, linguistic and acoustic interferences occur but when the masker speech is uttered in an unknown language, only acoustic interference is produced.

The idea that high-level linguistic knowledge serves to predict acoustic input is also supported by Billig, Davis, Deeks, Monstrey, and Carlyon (2013). In an ABA streaming task, listeners had to subjectively and constantly report if they heard one or two streams. The originality of the stimuli is that words were streamed into non-words and non-words into words. The overall result showed that auditory stream formation is influenced by linguistic processing and the build up effect is longer for words transformed into non-words than for non-words transformed into words.

In general, studies on phonemic restoration, informational masking and streaming are consistent and show that the perception of speech is not limited to acoustic cues *per se* and is deeply influenced by our knowledge of the language.

1.5 Rationale

All the research mentioned in the previous sections highlighted the interaction between the perceptive processes that separate two speech streams with cognitive functions, in particular attention, or our linguistic knowledge. Altogether, behavioural and neuropsychological studies attempt to understand the same cocktail party phenomenon whether it is under the scope of understanding speech stream separation or selective attention processes.

However, a shortcoming of these neurophysiological studies is that, with few exceptions (*e.g.* Akram, Presacco, Simon, Shamma, & Babadi, 2016; Akram, Simon, & Babadi, 2017; Miran et al., 2018), their analyses of the relationships between neural responses and the target speech envelope are based on an assumption: that listeners are able to maintain their attention focused on the attended speech stream for relatively long periods of time – at least a few tens of seconds and, in many of these studies, several tens of minutes. The main reason for using such long stimuli is that adequate decoding of neural responses requires a lot of data.

Studies suggest that the total duration of stimuli should be at least 15 minutes (Mirkovic, Debener, Jaeger, & Vos, 2015) and that the stimulus itself should last at least 10 seconds (O'Sullivan et al., 2015). Our own experience as participants in tasks involving such long stimuli, and informal reports from others, strongly suggest that this constant-attention assumption may not be warranted. Rather, it appears that for many listeners, keeping one's attention focused constantly for several tens of seconds on a speech stream, such as a voice telling a story, while another speech stream is being played concurrently, at approximately the same sound level, is a demanding task (*e.g.* Hoen et al., 2007; Shavit-Cohen & Zion Golumbic, 2019). Despite their best efforts, listeners are incapable of preventing momentary shifts in attention, from the target to the non-target voice.

In fact, when a behavioural approach is used, experimenters seem to favour relatively short stimuli to avoid this type of pitfall. Series of short, unrelated stimuli provide a reprieve for the participant who has an opportunity to restore their attentional resources before moving on to the next stimuli. While the sensory mechanisms in the perception of long and short stimuli may be largely identical, the cognitive mechanisms involved in the two situations may differ greatly. For instance, Conway et al. (2001) highlighted the importance of working memory in selective attention tasks. Processing short, sentence-like stimuli likely requires a very different involvement of working memory than long, story-like stimuli. In addition, semantic context can influence speech perception (*e.g.* Clarke et al., 2014). Shorter stimuli have naturally relatively limited semantic context compared to the long stories used in the neurophysiological studies mentioned above. Another effect, which may influence the dynamics of auditory selective attention relates to the notion of the build-up effect. This build-up can be more or less rapid, depending on how perceptually distinguishable the two streams are. Secondly, once the two concurrent streams are separated perceptually, the listener may need additional time to focus attention selectively onto the target stream. In addition, Billig et al. (2013) reported that the build-up can also be influenced by semantic information. In short, these rather slow phenomena are at risk of being missed by methods that involve only short stimuli.

These considerations highlight the importance of the duration of the stimuli, potentially leading to having different cognitive mechanisms involved, therefore making behavioural studies somewhat irreconcilable with a number of neurophysiological studies. On the one hand, behavioural studies use short stimuli, offering better control of the attentional focus of the participant, but are relatively limited

in the cognitive processes they might involve. On the other hand, neurophysiological studies, using long stimuli, likely involve more ecological cognitive processes, but offer little or no control of the attentional focus of the participant.

With the aim to address these limitations, a concurrent-speech intelligibility task — *Long-SWoRD* — was designed. A set of stimuli was assembled in order to infer fluctuations in auditory selective attention while participants are listening to short stories presented concurrently. In this context, the first goal of this manuscript was to document the performance of normal-hearing listeners in this task in situations where the perceptual separability of the competing voices ranges from easy to hard using a combination of voice and binaural cues. The second purpose is to bridge the gap between behavioural and neuropsychological studies and to observe if the results obtained using the two different approaches are consistent with each other.

This thesis is structured into four parts. In the first part, the new material and task are described as well as their use in an experiment to measure selective auditory attention to concurrent voices (see Chapter 2). In the second and third parts, performance in the separation of concurrent voices under adverse listening conditions is examined with a behavioural measure (see Chapter 3) and with neurophysiological measures (see Chapter 4). The fourth part explores the combination of behavioural and neurophysiological measures for modelling fluctuations in auditory selective attention (see Chapter 5). Finally, general conclusions and future developments are outlined in Chapter 6.

Chapter 2

The Long-SWoRD test

This chapter was partially published in:

Huet, M. P., Micheyl, C., Gaudrain, E., & Parizet, E. (2018). Who Are You Listening to? Towards a Dynamic Measure of Auditory Attention to Speech-on-speech. *Interspeech 2018*, 2272-2275.

2.1 Introduction

Assessing effective communication, and more particularly speech intelligibility, is an essential requirement for audiologist clinicians and researchers in psychoacoustic or communication sciences. Despite the recurring use of the term *"intelligibility"* in the literature, what is meant by this notion is still debated. As already mentioned in Chapter 1, intelligibility can cover two notions. The first notion covers the idea that a listener is able to retrieve a spoken message based only on the acoustic signal. The second notion is broader than the first one: a listener can be helped, in addition to acoustic signal, by any other verbal (*e.g.* syntax, semantics) or non-verbal (*e.g.* facial expression, gesture, *etc.*) sources to retrieve a spoken message. This latter notion is sometimes referred as *comprehension*.

In a review of intelligibility measures, N. Miller (2013) points out that the two main debates concern the intelligibility definition as well as the question of how the intelligibility should be measured. However, these two questions – definition

and measures – may not be independent and the definition of intelligibility may lie in the description of its measurement. In her essay on linguistics, Sadek-Khalil (1997) argues that, when a word is defined, the statement is about everything that word implies that is already known. In other words, the essence of intelligibility cannot be captured without taking into account the method used to evaluate it. Therefore, speech intelligibility should be understood as a *measure*.

Thus, in order to better understand the concept of intelligibility, different measures will be described in the next section. The first Section 2.1.1 will present measures that relate to the first concept of intelligibility mentioned above. The second Section 2.1.2 will cover measures that are associated with the second notion of comprehension.

2.1.1 Measures of intelligibility

A *word recognition test* is one approach, widely used in the clinical field, to measure intelligibility. Usually, a list of words is uttered and listeners report the word they believe they heard. The score of the subjects' performance can be computed either by counting the correct number of words or either by counting the number of correct phonemes. The phoneme scoring gives more sensitive results than the whole-word scoring since there are several phonemes in a word, which increases the scoring accuracy (Markides, 1978). The phoneme scoring usually yields scores that are on the order of 20% higher than scores for whole-words scoring (Olsen, Van Tasell, & Speaks, 1997). The scoring recommendations are different depending on the people tested. People who suffer from hearing loss or people who use hearing aids tend to have lower scores. To avoid floor effects, the use of phoneme scoring system is therefore recommended in this situation. Conversely, people with normal hearing tend to have scores that saturate upwards. A more restrictive scoring such as whole-words scoring is therefore more appropriate. In addition, the whole-words scoring is easier and faster. Lafon's test (Lafon, 1964) is an example of these lists. In Lafon's test, each word is phonetically balanced, which means that the frequency of the words in these lists match the frequency in natural language.

In the 1980s, tests for speech intelligibility with an automatic scoring were developed. The *Coordinate Response Measure (CRM)* is one of the most popular speech-on-speech tests and has been developed at first by Moore for the U.S. Air Force (Bolia, Nelson, Ericson, & Simpson, 2000; T. Moore, 1981). This

speech corpus consists of a call sign with a colour-number combination. The 256 sentences of this test are a factorial combination of 8 call signs (*"arrow", "baron","charlie","eagle", "hopper", "laker", "ringo", "tiger"*), 4 colours (*"blue", "green", "red,"white"*) and the numbers between 1 and 8. For instance, *"Ready baron, go to the red one now"* is a typical sentence where *"baron"* is the call sign and *"red one"* is the colour-number combination. At the beginning of the task, the listener is assigned with a call sign and has to answer the correct colour-number combination. The percentage of correct number and colour identification is usually used as a measure of speech intelligibility (Brungart, 2001).

During the same period, the first Matrix test was developed (Hagerman, 1982) for Swedish. Based on the same principle as CRM, Matrix test sentences have a fixed structure "name-verb-numeral-adjective-noun". The Matrix material consists of 10 first names, 10 verbs, 10 numerals, 10 adjectives and 10 nouns for a total of 100 000 different sentences (see Table 2.1). In the past few years, the organization *HörTech GmbH* has extended the International Matrix test, also known as *the Oldenburger Satztest (OLSA)* in several languages (for a review, see Kollmeier et al., 2015).

Table 2.1 International Matrix test

Name	**Verb**	**Number**	**Adjective**	**Noun**
Peter	got	three	large	desks
Kathy	sees	nine	small	chairs
Lucy	brought	seven	old	tables
Alan	gives	eight	dark	toys
Rachel	sold	four	heavy	spoons
William	prefers	nineteen	green	windows
Steven	has	two	chap	sofas
Thomas	kept	fifteen	pretty	rings
Doris	ordered	twelve	red	flowers
Nina	wants	sixty	white	houses

The CRM and the Matrix test are typical examples of *closed-set tests*. They have a fixed grammatical structure, an unpredictable semantic content and a limited answer set. According to Nilsson, Soli, and Sullivan (1994), the small size of these corpora could lead to learning the material and could therefore be a potential barrier to the use of these limited corpora. It is for this reason that the latter authors developed the *Hearing In Noise Test (HINT)*. This test consists of 12 lists of 20 short sentences. Unlike the CRM or the Matrix test, the answers

are not suggested to subjects who must repeat as best as they can what they have heard. The rate of correct answers is a scoring on a word-by-word basis. On the same idea and more recently, Helfer and Freyman (2009) developed the *Theo, Victor or Michael (TVM)* sentences test. Each sentence has the structure " Call sign *discussed the* ___ *and the* ___ *today"*, where the Call sign is a name (Theo, Victor or Michael) and blanks are words that participants have to repeat. The HINT and the TVM are examples of *open-set tests*. They have a semantic close to everyday sentences and a more flexible grammatical structure. Corpora such as the Harvard/IEEE (IEEE Audio and Electroacoustics Group, 1969) can also be mentioned. This corpus is composed of 72 lists of 10 sentences. Five keywords are embedded in each rich, meaningful sentence such as *"These days a chicken leg is a rare dish"*. The IEEE corpus is among the oldest and most popular tests; however, the level of complexity of its language makes it unsuitable for people who are hearing impaired or have memory deficits (Sharma, Tripathy, & Saxena, 2016).

In general, closed-set and open-set tests differ in terms of stimulus diversity (limited for closed-set tests *vs.* varied for open-set tests) and in terms of answer type (multiple choice for the closed-set tests *vs.* repetition for open-set tests). A few studies (Clopper, Pisoni, & Tierney, 2006; Yu & Schlauch, 2019) compare both test types. According to these studies, open-set tests are more precise and more difficult than closed-set tasks. According to Clopper et al. (2006), the difficulty of the open-set tests might come from the comparison listeners have to do between the stimulus item and all the words they have in their lexicon whereas the comparison in closed-set involves fewer words.

2.1.2 Measures of comprehension

One shortcoming is that all the speech intelligibility measures aforementioned tend to be focused on relatively simple listening situations. Complex listening environments, by contrast, involve the extraction of meaning, and, generally, the formulation of a valid reply.There are mainly two ways to assess how good listeners are at a *comprehension* task: questionnaires and comprehension tests. One is to ask listeners to answer a series of content-related questions about what was heard. For instance, the questionnaires used to measure the behaviour in the neurophysiological studies, mentioned in Chapter 5, are based on this approach (*e.g.* Ding & Simon, 2012b; O'Sullivan et al., 2015). This testing format might

introduce long stimuli and hence a significant memory requirement which may interfere with the comprehension measure. Some questionnaires use an "on-the-go" approach which means to query the participant during the stimulus instead of at the end of the presentation (*e.g.* Best, Keidser, Buchholz, & Freeston, 2016). An alternative approach used to bypass the memory load is to reduce stimuli to brief question-answer pairs (*e.g.* Best, Streeter, Roverud, Mason, & Kidd, 2016). Finally, to ensure that answers are based on semantic information rather than just keywords, some questionnaires only use synonyms as possible answers (*e.g.* Hafter, Xia, & Kalluri, 2013).

According to Hustad (2008), a limitation of these questionnaires is that they are distant from real communication where a speaker is talking to a conversation partner and the partner must respond or react to the speaker. Comprehension tests address this issue by assessing the ability of participants to execute a requested action on a correct object. Fontan, Tardieu, Gaillard, Woisard, and Ruiz (2015) examined the relationship between speech intelligibility (assessed with a word recognition tests) and speech comprehension (assessed with a comprehension test on computer) in speech-on-speech. The two tests used identical stimuli but differed in the way participants answered. Listeners had to repeat as best as they could what they heard in the intelligibility task, whereas they had to perform the command in the comprehension task. The main outcome is that intelligibly cannot fully predict listeners' comprehension. Theses two measures appear to provide different insights and, according to the authors, should be used complementarily. However, it should be noted that the results may also be difficult to compare due to differences in answer protocol. In the comprehension test, participants had the objects to be manipulated in front of them, whereas in the intelligibility test, subjects had to repeat without any outside help.

2.2 Rationale for a new test

As mentioned in section 2.1, intelligibility can cover multiple concepts: the intelligibility of a spoken message or the broader comprehension of this spoken message. The descriptions of the measures also provided a better understanding of what intelligibility is in general. The literature that focuses on the first notion of intelligibility, retrieving a spoken message based on a sound signal, is much more prevalent in speech-in-noise studies and more particularly in speech-on-speech

studies. This is why the rest of this chapter will focus more particularly on this simple notion of intelligibility.

As mentioned before, in measures of intelligibility, tests are divided into two categories – closed-set tests and open-set tests – each of which has its strengths. The advantage of closed-set test lies in the easy scoring process applied to participants' responses while open-set tests allow more flexibility in variety and extent of stimuli. We have therefore created an intelligibility test which combines the two advantages of closed-set tests and open-set tests: long stimuli and objective scoring. This Selective Word Recognition Discrimination (SWoRD) test will be used with Long stimuli and will be named: the *Long-SWoRD* test.

As also mentioned in Chapter 1, it appears that for many listeners, keeping one's attention focused constantly for several tens of seconds on a speech stream, such as a voice telling a story, while another speech stream is being played concurrently, is a demanding task (*e.g.* Hoen et al., 2007; Shavit-Cohen & Zion Golumbic, 2019). As a result, listeners might be incapable of preventing momentary shifts in attention, from the target to the non-target voice. It is therefore important to be able to infer which speaker participants are listening to at different points in time. Thus, instead of asking the participant to retrieve a single word, they are instructed to find three words from different key time points (or keywords) belonging to the story they have to listen to.

The development of two versions of the Long-SWoRD test will be presented in the next sections.

2.3 The Long-SWoRD v1: development of the test materials

2.3.1 Selection of speech material

The main criterion for the stimuli selection was an interesting semantic content, so that participants would pay attention to it. The French book *Le Charme discret de l'intestin* [Gut: the inside story of our body's most underrated organ] (Enders, Enders, & Liber, 2015) describes the role that our intestine, this "second brain", and its microbiota play in health. It is an informative book compatible

with the selection of meaningful and engaging stories. In addition, the French Audiobook published by *Audiolib* (Enders, Monceau, & Liber, 2016) is narrated by a speaker's voice which can be credibly modified by STRAIGHT (see Chapter 3). 547 anecdotes and fun facts were extracted from the audiobook according to several criteria such as the time (11 - 18 seconds) and the number of words per story (22 - 55). Below are two examples of stories. The first story has a somewhat common semantic context while the second story has a semantic context oriented around the topic of digestion, the subject of the book.

1. *"Nous racontons parfois à nos enfants des mensonges plus gros qu'eux. Je pense par exemple au mensonge du bonhomme barbu qui, une fois par an, pointe son nez et sa hotte pour offrir des cadeaux aux enfants avant de repartir sur un véhicule à la croisée du tapis volant et de la charrette à bœufs."* [1]

2. *"Un morceau d'entrecôte peut par exemple se balancer six heures dans notre petit hamac avant d'être livré intégralement à l'intestin grêle. Pas étonnant, donc, que nous ayons une terrible envie de dessert après avoir mangé de la viande ou des beignets bien gras."* [2]

Then, the keywords (words that the subjects had to retrieve in the stories) were selected based on two criteria: their position in the story and their frequency in the language. Hence, three keywords were selected at different key times in the story: a keyword at the beginning of the story, a keyword in the middle and a keyword at the end. However, the very first and last words of the stories were excluded from the selection in order to minimize the effects of recency and primacy. In addition, the selected keywords could only appear one time in the story to avoid repetition. All the selected keywords were compared to a lexical database, "Lexique" (New, Pallier, Ferrand, & Matos, 2001), containing the occurrences of all these words in the French language. Words that did not appear in the database as well as words that were too rare or too frequent were replaced with another word from the sentence or deleted from the keyword set. As a result, if one word was deleted, the entire story was removed, even if the other two words met the

[1] Translation: "Sometimes we tell fibs to our children. We do it because these little untruths are so nice. There's the one about the man with the big white beard who arrives once a year on his tuned-up reindeer sleigh piled high with children's presents."(Enders, Enders, & Shaw, 2015)

[2] Translation: "A piece of steak may easily be churned about for six hours before all of it has disappeared into the small intestine. This explains why we often fancy a sweet dessert after eating meat or fatty, fried foods."(Enders, Enders, & Shaw, 2015)

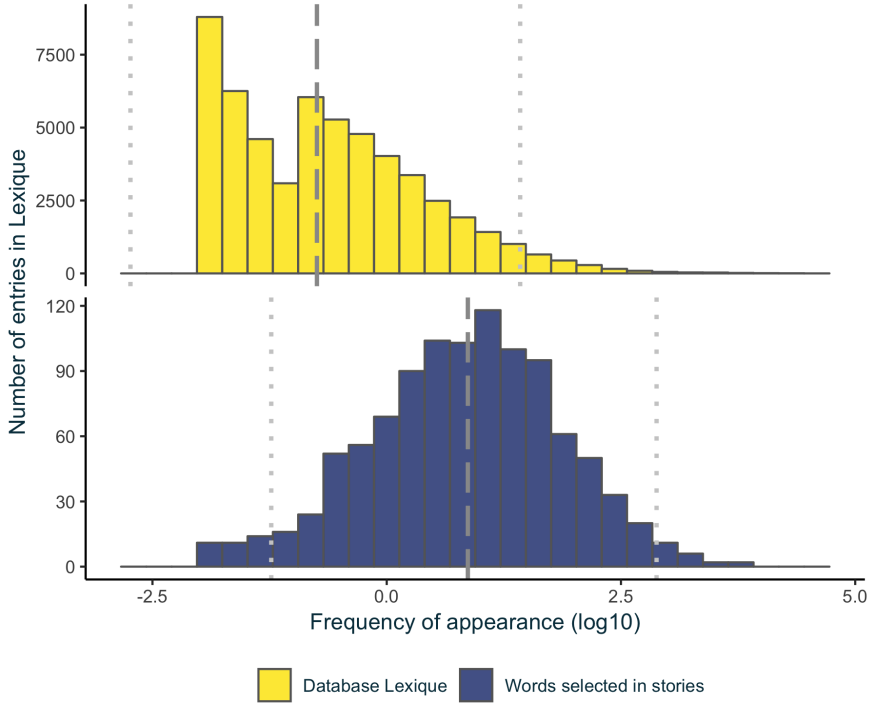

Figure 2.1 : Words distributions in "Lexique" (New et al., 2001) (in yellow) and the Long-SWoRD keywords set (in blue). The vertical lines represent the mean and 2 standard deviations for each distribution.

selection criteria. These modifications reduced the total number of stories from 547 to 526. Figure 2.1 shows the words distribution in both the lexical database and the keywords set. Finally, the selected keyword distribution had a geometrical mean of 6.67 per million occurrences in the French language and 95% of values range between 0.04 and 532.79 per million occurrences.

The next step of the test selection was to match the target and masker stories. To form a pair, the two stories had to be of the same duration and they could not contain the three keywords selected in the other story. At this point, the material was composed of trials including a target story and a masker story as well as their six associated keywords. In order to make the task more challenging, three words that do not appear in either story have been assigned to each pair. These *extraneous* keywords originated from another story. All 9 different keywords (3 target, 3 masker and 3 extraneous) were presented visually together.

2.3.2 Evaluation of the test

Online study

Detecting three keywords in stories that could sometimes last 18 seconds seemed potentially complicated for the participants because of the cognitive load of the working memory. It was therefore decided to evaluate this experimental procedure with an online study. 231 participants took part in that study but 12 participants were removed from the analysis because French was not their native language. The average age of the remaining 219 participants is 33.92 years old ($\sigma = 14.26\%$). The task was to find the 3 keywords of 20 stories presented in isolation condition (without masker).

Each participant was exposed to 20 stories chosen at random from the full set. As a result, the total number of presentations for each story varies between 1 and 16. The participants' average score was 89.45% ($\sigma = 9.74\%$). Within each story, the distribution of the score for each word was wider ($\sigma = 17.1\%$).

From the 263 pairs of stories, only 178 were kept on the basis of three criteria. First, both target and masker stories had to have been tested at least three times each. Second, the average score of both target and masker stories were superior to 69.86% ($89.45\% - 2 \times 9.74\%$). Third, neither the keywords from the target story nor the keywords from the masker story had an individual score inferior to 55.14% ($89.45\% - 2 \times 17.1\%$). Thus, stories that were not tested enough or too difficult in an isolation condition were removed from the corpus.

Regrouping into lists

166 stories were selected and divided into 12 lists. The condition to regroup the stories was that none of the 9 keywords associated with a pair could appear within the other stories of the list. Finally, 12 stories per list were selected in order to obtain the same duration from one list to another. In the end, the material consisted of 12 cleaned up lists of story pairs counterbalanced in duration. All lists had a similar total duration ($\mu = 175.08$ seconds, $\sigma = 0.9$) and within each list, there were equally short and long stories (see Appendix B for the material).

2.3.3 Extraneous keywords analysis

Word2vec

The extraneous keywords of a story are target and masker keywords originating from other stories. Since all the stories come from the same book and by extension from a similar lexical field, it was necessary to ensure that the extraneous keywords were not semantically closer to target or to the masker keywords. To ensure that extraneous keywords have as little influence as possible on the participants' choice, the semantic similarity was measured between the target, the masker and the extraneous keywords.

It is not easy to determine how close two words are semantically, but Mikolov, Chen, Corrado, and Dean (2013) from Google have developed a new technique to estimate this measure called *word2vec*. The main idea is to represent a word in vector space with a simple neural network trained with a single hidden layer. This neural network is fed with word pairs found in a training *corpus*. The goal is to use the hidden layer weights as "word vectors". The below example shows a small neural network trained on this text composed with four sentences:

"There are eight *planets* in the *Solar System*. The sixth *planet* is *Saturn*. *Saturn* is *famous* for its *rings*. The *rings* of *Saturn* are the most *extensive ring system*"

For each sentence, the neural network is fed with nearby words. Table 2.2 shows the word pairs used to train this example and figure 2.2 represents the neural network. At the end, word vectors are positioned in the vector space such that words sharing common semantic contexts in the corpus are located close to each other in that space. For instance, according to this neural network, the word "Saturn" has a high probability to appear nearby the word "Ring".

Table 2.2 Word pairs for word2vec example.

Sentence 1	**Sentence 2**	**Sentence 3**	**Sentence 4**
(Planet, Solar) (Planet, System) (Solar, System)	(Saturn, Planet)	(Saturn, Famous) (Saturn, Ring) (Famous, Ring)	(Ring, Saturn) (Ring, Extensive) (Saturn, Extensive) (Saturn, Ring) (Extensive, Ring) (Extensive, System) (Ring, System)

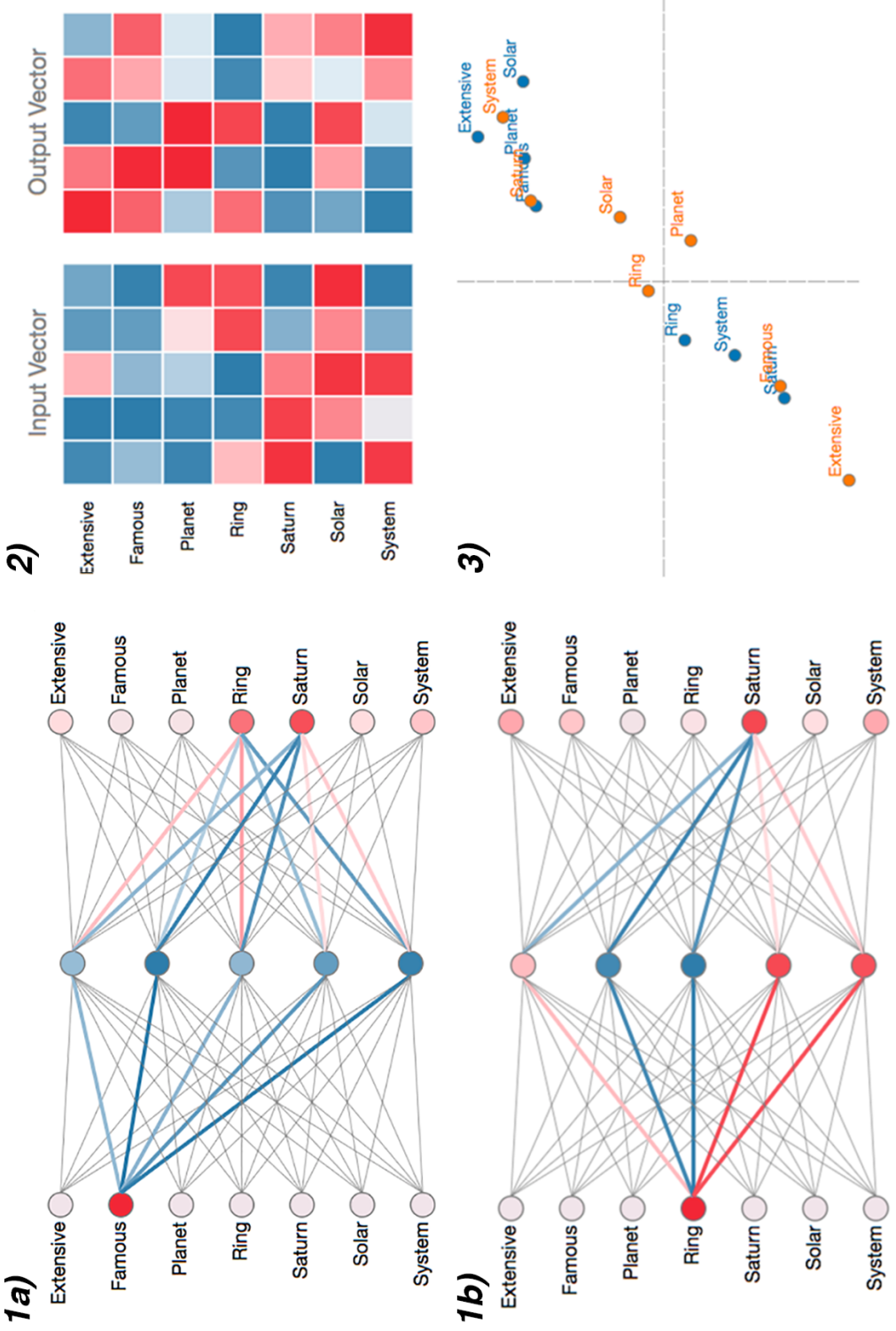

Figure 2.2 : Neural network is represented for the input word "Famous" and for the input word "Ring" for 1. The positive active levels of neurons are in red and the negative levels of neurons are in blue. In 1a example, the words "Saturn" and "Ring" have high probabilities to appear nearby the word "Famous". In 1b example, the words "Saturn", "Extensive" and "System" have high probabilities to appear nearby the word "Ring". The word "Saturn" has the highest probability (in red). 2 shows the weighted word vectors. Input word vector (in blue) and output word vector (in orange) are also represented in the space vector 3. (Adapted from Rong, 2016).

Semantic similarity between keywords

A French word2vec model was used (Gaudrain & Crouzet, 2019) to estimate the semantic similarity between the target, the masker and the external keywords for each story pair. This model was trained with the French Wikipedia corpus. Figure 2.3 shows the semantic similarities between target and masker keywords, between target and extraneous keywords and between masker and extraneous keywords. The similarities vary from 0 (not semantic similar) to 1 (semantic similar) and there is no difference between the target-masker similarity ($\mu = 0.227$), the target-extraneous similarity ($\mu = 0.229$) and the masker-extraneous similarity ($\mu = 0.228$) [$F(2, 286) = 0.12, p = .88$]. Therefore, extraneous keywords are as close to target keywords as masker keywords, from the point of view of meaning and should not overly influence participants.

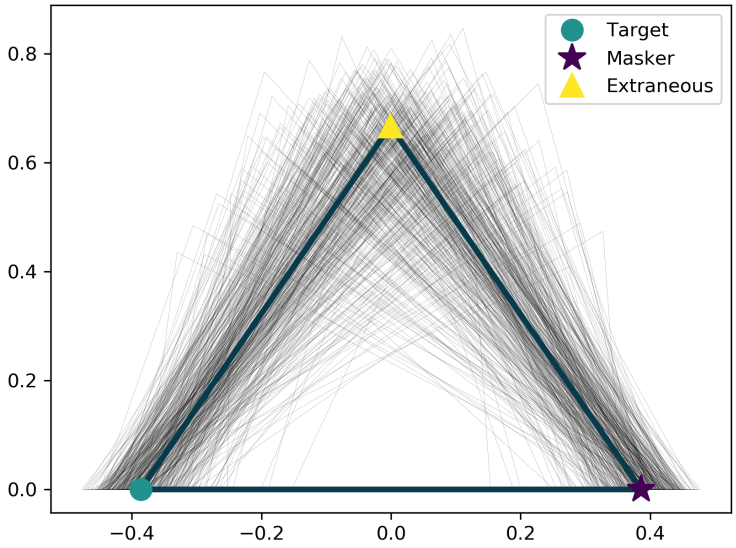

Figure 2.3 : Semantic similarities between the target (circle green), the masker (purple star) and the extraneous (yellow triangle) words. The similarities for each story pair and for each word position (i.e. Beginning, middle and end of story) are represented with grey lines. The dark blue line shows the average similarity. For illustration purpose, all the lines are (1-similarity).

2.3.4 Final version

The final version of Long-SWoRD v1 therefore contains 12 lists composed of 12 stories (see Appendix B for the material). Figure 2.4 represents three target and three masker keywords within each story. Due to the many constraints mentioned below to select keywords, it can happen that in some trials the masker keywords overlap the target keywords and in other trials, the maskers and target keywords are spaced a few seconds apart. The 9 keywords associated with the trial are then presented to the participant in the form of a 3×3 matrix (see Table 2.3). In order to avoid that the participants answer both masker and target associated with the same position (e.g. the target keyword "fondamentaux" and the masker keyword "trouble"), which would make it impossible to infer which story the participant is listening to, only one answer per line is possible. Thus, each line of the matrix includes one target, one masker and one extraneous keyword.

Table 2.3 Answers for a trial. Target keywords are in green, maskers keywords in violet and extraneous keywords in black.

expérience	**fondamentaux**	**trouble**
sévère	participants	**vue**
chemin	comportement	**héréditaire**

2.4 The Long-SWoRD v2: additional material

A second version of the Long-SWoRD test has been developed to allow a larger distribution of the material, the first version being subject to copyright related to the book by Giulia Enders and the audiobook publishing house Audiolib. This second version is not finalized but has already been used in an experiment about rhythmic priming and language. This section presents the work already done on the second version of the Long-SWoRD test. The material used in the rhythmic priming experiment is presented in Appendix C.

2.4.1 Creation and recording of the speech material

700 stories were created by speech therapy students. The main criterion for the stories creation was identical to the first version: the story had to be interesting so that participants would pay attention to it. Below is an example of a story:

Figure 2.4 : Final procedure of the Long-SWoRD test with the target story [a] above (in blue) and the masker story [b] below (in green-blue). Target keywords (green) and masker keywords (violet) are scattered throughout both stories.

[a] Translation: "Smell is one of our most *basic* senses. Unlike taste, hearing, or *vision*, smells are not checked out before they make their *way* to our consciousness." (Enders, Enders, & Shaw, 2015)

[b] Translation: "As with lactose intolerance, this intestinal functional *disorder* also exists in a *severe* congenital form: fructosemia or *hereditary* fructose intolerance."

"Au football américain, la plupart des joueurs ont de grandes traces noires placées sous les yeux. Leur rôle premier est non pas de leur donner un côté guerrier, mais de diminuer la réflexion de la lumière des projecteurs du stade ou du soleil. Cela évite aux joueurs d'être aveuglés." [3]

Then, these 700 stories were recorded by a female speaker. The stories were naturally spoken and there were no grammatical errors. The duration of the stories varies from 10 to 25 seconds.

2.4.2 Keywords

This second version keywords criteria have been selected according to the same criteria that the keywords of the first version. Hence, three keywords were selected at different key times in the story (beginning, middle and end of the story). The very first and last words of the stories, as well as the words that appear more than one time, were excluded. For the above example, the words *"traces"*, *"guerrier"* and *"évite"* were selected.

The choice of external words is different from the first version. Instead of using keywords originating from other stories, we selected synonyms and antonyms. The main reason for this change in extraneous keywords is that the task seems too easy in the first version of Long-SWoRD (see Chapter 3). A higher number of more distracting external words will make this second version of the test more difficult for participants.

2.5 Conclusion

The two versions of the Long-SWoRD test were presented in the section 2.3 and section 2.4 respectively. However, as the second version of the test is not yet finalized, the next chapters will focus exclusively on version 1. Behavioural studies with the Long-SWoRD test will be presented in Chapter 3 and a neurophysiological study in Chapter 4.

[3] Translation: "In American football, most players have large black marks under their eyes. Their primary role is not make them look like warriors, but to reduce the reflection of the light from the stadium spotlights or the sun. This prevents players from being blinded."

Chapter 3

Talker segregation with the Long-SWoRD test

This chapter was partially published in:

Huet, M. P., Micheyl, C., Gaudrain, E., & Parizet, E. (2018). Who Are You Listening to? Towards a Dynamic Measure of Auditory Attention to Speech-on-speech. *Interspeech 2018*, 2272-2275.

3.1 Introduction

Being able to follow a conversation among competing speakers is perhaps one of the most important functions of the sense of hearing in humans. In the scenario where two speakers compete for your attention, *(speech-on-speech)*, to successfully negotiate the challenge of listening to one speech stream while ignoring the other, one needs to properly separate the voice they are trying to attend *(target stream)* from the one they are trying to ignore *(the masker stream)*. As already mentioned in Chapter 1, this difficult task is facilitated by different cues. Bregman (1994) classifies these cues, used for grouping and segregation, into two main categories, namely the *"primitive"* and the *"schema-based"* grouping.

According to Bregman (1994), the primitive process is based on grouping/separating elements that share a common acoustical feature. This process

is driven by the stimulus itself and does not rely on knowledge. The cues are automatically extracted and processed by the primary auditory pathway. This process is referred to as a *bottom-up* process. On the other hand, the schema-based grouping is based on learnt expectations such as knowledge of familiar sounds, acquired concepts of grammar. This descending process is referred to as a *top-down* process. However, these two segregation processes should not be considered as independent. They are likely to collaborate and adapt to each other.

3.1.1 Primitive grouping

According to Bronkhorst (2015), two main primitives cues help listeners to separate two speakers and have been extensively studied. These two cues, namely the spatialisation of the speakers and the characteristics of the voices, will be briefly presented in the next paragraphs.

Spatialisation

Cherry's seminal article (1953) continues to be frequently cited for highlighting the importance of the factors governing communication performance in a *"cocktail party"* environment. Among those factors, the spatial separation of sound sources subsequently received the greatest attention in the literature. Cherry (1953) presented two spoken sentences dichotically — one speech stream to each ear — and noted that recognition of one sentence improved compared to a condition in which the messages were presented diotically — both speech streams to both ears — both mixed together on a tape.

In a dichotic stimuli presentation, Cherry (1953) also observed that participants are generally able to report almost nothing about the content of the unattended message. Participants even frequently failed to notice a change from English to German in the unattended channel. Despite these results, it seems that the unattended information is still processed to some extent and that keywords can divert attention toward the unattended stream, such as the listener's own name (Moray, 1959).

The dichotic and diotic presentations cannot reflect how human's perceptual system work in real life since all sound sources are processed binaurally (with two ears) for normal hearing people. The auditory system also uses the tiny difference in time of arrival between the two ears or the interaural time differences

(ITDs) and the sound intensity or the interaural level differences (ILDs) to locate the source of a sound (for more information and reviews, see Bronkhorst, 2015; Middlebrooks, 2017). More generally, stimuli are convolved through head-related transfer functions (HRTFs) that result from sounds experiencing spectral modifications during their propagation through and around the head.

As a result, instead of using dichotic and diotic listening conditions, presenting targets and masker speech streams from different speakers at different angles relative to the participant is one of the most direct approaches to study spatial groupings. In their report, Ericson and McKinley (2001) analysed participants' performance in a CRM task when two speakers were presented dichotically and diotically. They additionally presented the two speaker with different angles relative to the participants by manipulating the mixing of the recordings in a headphone. They found that the dichotic presentation of the competing speakers was always more intelligible than the diotic presentation or than an angle of 45° presentation. In addition, participants' performance was maximized with a angle of 90° presentation and further separation did not yield higher performance. A few years later, Brungart and Simpson (2007) found that, in a CRM task, two speakers can be perceptually segregated as long as they are physically separated by an angle of 10°. Helfer and Freyman (2009) also found an effect of stimuli spatialisation with the open-set TVM corpus sentences. However, the study of spatial grouping of speech streams is complicated by the fact that spatial separation cues are mixed with other separation cues such as the spectral envelope or the F0 (see next section).

Vocal characteristics

When both speakers are physically close to each other (less than an angle of 10°) or when they are mixed together into one single channel as it might happen in real life with headphones or with the phone, the challenge of separating the voices becomes more complicated. Despite some evidence that differences in the levels of the two voices, also known as the *target-to-masker ratio (TMR)*, can help listeners to perform that difficult task (e.g. Brungart, Simpson, Ericson, & Scott, 2001; Egan, Carterette, & Thwing, 1954), the vocal characteristics of the competing talkers seem to be the most powerful monaural speech stream separation cues (Bregman, 1994; Brungart et al., 2001). To discriminate two voices at the same level (at a TMR of 0 dB), the apparent gender of the voice seems to be one of the most

helpful cues, as pointed out by Brungart et al. (2001). In their study, Brungart et al. (2001) found that participants had better performances in the CRM task when the target sentence and the masking sentence were spoken by talkers of different sex than when target and masker sentences were spoken by talkers of the same sex.

Figure 3.1 represents how two physical characteristics, namely the fundamental frequency (F0) and the vocal tract length (VTL), can be manipulated to change the apparent gender of a synthesized speaker (Skuk Verena G. & Schweinberger Stefan R., 2014). F0 is the rate of the glottal pulses of air in the vocal tract produced when human voice utter vowel or sonorant consonants and is perceived by the ear as pitch. The averages of fundamental frequencies of vowels vary between 210 and 235 Hz for women and between 124 and 141 Hz for men (Peterson & Barney, 1952). Regarding the VTL acoustics cue, resonances responsible for the formant frequencies in the spectral envelope are constrained by the length of the vocal tract (Chiba & Kajiyama, 1941; Fant, 1970). Male vocal tracts being longer than women's vocal tracts, men's formant frequencies are around 16% lower than women (Peterson & Barney, 1952).

Darwin, Brungart, and Simpson (2003) studied, in three experiments, how F0 and VTL affect performance on a CRM task. The first experiment only studied how increases in the natural F0 difference between two CRM sentences affected listeners' ability to attend to the target story. Systematic improvements in performance were produced with differences greater than 2.0 semitones (st) in F0. But the level of performance found with different-sex talkers cannot be reached with F0 alone. The second experiment studied how VTL only can affect participants' performance to attend to the target story. Systematic improvements in performances were produced when the ratio of lengths was 1.08 (1.3 st) or greater but similar to the first experience, the level of performance found with different-sex talkers cannot be reached with VTL alone. Their third experiment used both F0 and VTL differences. An incremental shift in gender, simulated by systematic changes in both F0 and VTL, improved the participants' performance better than did differences in F0 or VTL alone. Finally, Darwin et al. (2003) noticed that performances were better when shifting one of two utterances spoken by a female voice towards a male voice than shifting male towards female.

Vestergaard, Fyson, and Patterson (2009) investigated how F0 and VTL interact when syllables are used as targets and maskers. When the TMR was 0 dB, participants' performance depended on the combination of F0 and VTL.

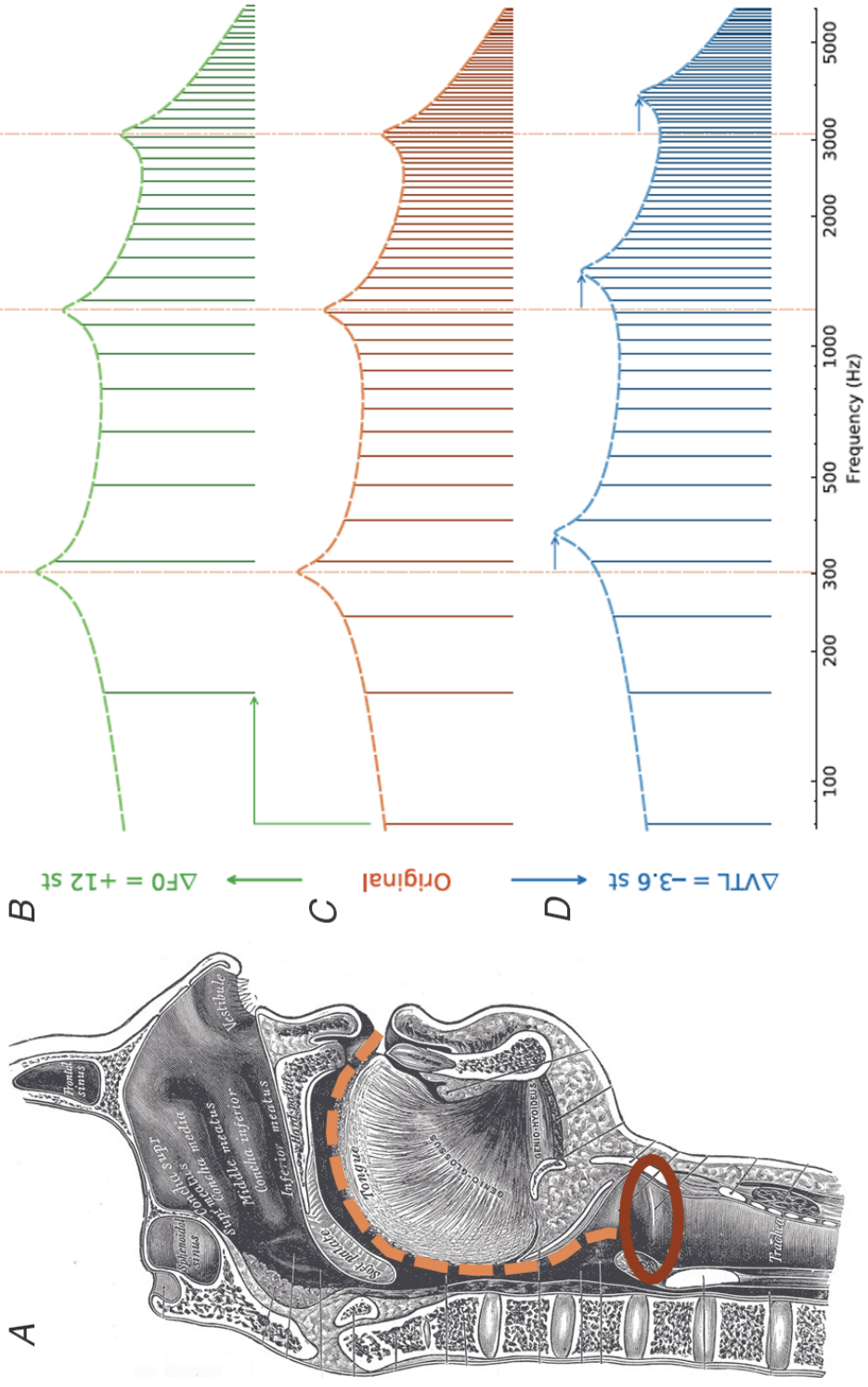

Figure 3.1 ***(A)*** *Sagittal section through a human head and neck. The glottal folds, circled in red, are the source of the sound of which F0 is a property. The dashed line represents the vocal tract from glottal folds to lips.* ***(B)****,* ***(C)*** *and* ***(D)*** *are magnitude spectra, that are, the distribution of energy across frequency and are composed of harmonics (vertical lines) and vocal-tract resonances (dashed line).* ***(C)*** *represented an idealized vowel.* ***(B)*** *shows the effect of increasing F0 by 12 st and* ***(D)*** *shows the effect of decreasing the vocal-tract length by 3.6 st (adapted from Gaudrain and Başkent (2018))*

When the characteristics of the target and the masker voices were largely different, performances was primarily determined by TMR.

Spatial and voice cues

Extending the work of Vestergaard, Fyson, and Patterson (2009), Ives, Vestergaard, Kistler, and Patterson (2010) showed that voice information and spatial location can be extracted for syllables. For spatialisation angles larger than 8° between the target and the masker, performance peaked and the effect of the vocal characteristics was obscured. Figure 3.2 shows how Vestergaard, Ives, and Patterson (2009) summarized their results. Ives et al. (2010) concluded that the benefit from the spatial cues was so large that no additional improvement from other cues such as vocal characteristics could be obtained.

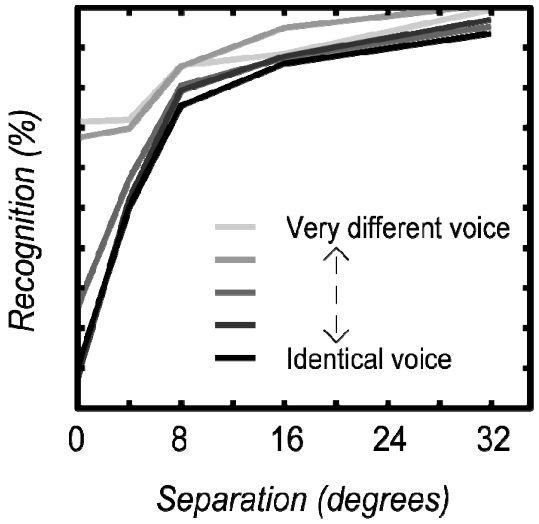

Figure 3.2 Speech recognition performances for five different voices used as a function of spatial separation (adapted from Vestergaard, Ives, & Patterson, 2009).

3.1.2 Schema-based grouping

Although primitive cues are clearly potent, they are not fully representative of real-life listening conditions. Two types of *a priori* knowledge can influence the way we separate two speech streams, namely our linguistic knowledge and *a priori* information about speakers.

Linguistic knowledge

When humans use language, expectations about upcoming material are created. These expectations play a role for language comprehension and are considered to result from predictions that are based on contextual information (semantic) and knowledge of the language (syntax) (for a review, see Kuperberg & Jaeger, 2016).

Several experiments have investigated whether two speech streams can be separated by using the linguistic contents of the target and the masker. Freyman, Balakrishnan, and Helfer (2001) used native and non-native talkers, time-reversed speech, or spoken in a foreign language speech streams as stimuli. Iyer, Brungart, and Simpson (2010) used CRM sentences, native, foreign or time-reversed speech stream as stimuli. Performance was relatively better when the masker is a foreign speech stream or a non-native language than when the masker is normal or time-reversed speech stream. Interestingly, the worst performance was for the CRM maskers that shared the same syntax and timing as the CRM target sentences and thus maximized confusions. These results support the idea that linguistic cues, such as syntax and semantics, can be used to separate concurrent speech streams. Freyman et al. (2001) study also showed that speech streams separation was always facilitated by the introduction of some spatial separation.

The explanation of these results might lie in the linguistic similarity hypothesis articulated by Brouwer, Van Engen, Calandruccio, and Bradlow (2012). In this latter study, they examined how the degree of similarity between target and masker streams impacted the segregation. The background speech languages (English *vs.* Dutch), the semantic content of the background speech (meaningful *vs.* meaningless sentences), and the listener status (native, non-native, or unfamiliar) were investigated. Results showed that the more similar the target stream is to the masker stream, the harder it is to efficiently segregate the two sentences effectively.

Other aspects of language have been studied such as accent (Calandruccio, Dhar, & Bradlow, 2010), phonetic distance between languages (Calandruccio, Brouwer, Van Engen, Dhar, & Bradlow, 2013) or dialect (Brouwer, 2017) and also support the linguistic similarity hypothesis.

Talkers' information

Helfer and Freyman (2009) extended Freyman et al.'s studies Freyman et al. and compared the relative efficacy of a semantic cue to a voice cue to direct the listener's attention to a target sentence using the open-set TVM sentences. In the semantic condition, the name of the person the participants had to listen to was projected on the screen. Then, the first word of each sentence was a name. Participants had to listen to the sentence that matches the name appearing on the screen. In the voice condition (referred as the *indexical* cue by the authors), each sentence started with the sentence *"This is the target voice"* uttered by the target stream. The obtained results demonstrated that both semantic and indexical cues can be equally used to direct listeners' attention on a target voice. This study therefore suggests that the knowledge that listeners have of the speaker offers them an advantage in a cocktail party situation.

These results are consistent with Newman and Evers (2007)'s study proposing that auditory stream segregation can be enhanced when participants were familiarized with the target voice. Johnsrude et al. (2013) investigated how the familiarity with spouse's voice could help participants to segregate two talkers in a CRM task. A benefit was observed when the familiar voice was used as a target and also as a masker. More recently, Holmes, Domingo, and Johnsrude (2018) showed that the advantage of a familiar voice in speech-on-speech tasks is maintained even when the acoustic cues (F0 and VTL) of the familiar voice were manipulated and hence, participants could not explicitly recognize the voice.

3.2 Rationale

As shown in section 3.1, many behavioural studies have investigated how two talkers are separated. The stimuli used in these studies are varied and range from simple syllables to whole sentences. As already mentioned in Chapter 1 and Chapter 2, short stimuli might underestimate the contribution of perceptual mechanisms such as the build-up effect or participants' knowledge of language. Stimuli from the Long-SWoRD, longer than those used in all the behavioural speech-on-speech studies, allow to reach a more real-life listening condition. Then the question of the participants' performance is raised. Will participants behave in a similar way to previous studies with the Long-SWoRD test? Or on the contrary, might the contribution of these perceptual mechanisms modify participants' performance?

In this chapter, two primitive segregation cues – the vocal characteristics and the spatialisation of the stimuli – are studied with the Long-SWoRD test in two experiments. These cues have been extensively studied and are clearly potent to disentangle target from masker speech streams. As previously observed, studies using nonsense sentences have shown semantic cues are powerful discriminating elements to segregate two speech streams (Freyman et al., 2001; Iyer et al., 2010). Hence, a specific schema-based cue, the linguistic knowledge of the participants, is also examined *a posteriori* in both experiments.

The second part of this chapter will describe the general methods common to both experiments. Descriptions of the measurement of the distance between the target voice and the masker voice as well as the measurement of the semantic context are detailed. The methods specific to each experiment, their results and conclusions are presented in section 3.4 (Experiment 1) and section 3.5 (Experiment 2) of this chapter, respectively. Finally, a general conclusion for this chapter will be presented in the last section.

3.3 General methods

3.3.1 Apparatus

Stimuli were presented with OpenSesame (Mathôt, Schreij, & Theeuwes, 2012). Participants listened to stimuli over a Sennheiser HD250 Linear II headphone in a sound-attenuated booth.

3.3.2 Stimuli

The *Long-SWoRD test* coupled with the audio stimuli from the audiobook *Le Charme discret de l'intestin* [The Inside Story of Our Body's Most Underrated Organ] (Enders et al., 2016) was exploited to perform experiments 1 and 2. This material was presented in Chapter 2.

3.3.3 Voice and manipulation

In studies with real speakers, it is difficult to determine how close or similar two voices are to each other (for a review, see Kreiman & Sidtis, 2011). In

general, two problems are encountered. First, there does not seem to be a consensus on the elements to be taken into account in determining the similarity between two voices, although the measurement of F0 and VTL seems ubiquitous. Secondly, the complexity of the elements that composed the voice makes this measure complicated. For example, estimating a speaker's VTL remains unreliable, although different approaches have been proposed (*e.g.* Flego, 2018; Turner, Walters, Monaghan, & Patterson, 2009).

Hence, the approach used in previous studies (e.g. Başkent & Gaudrain, 2016; Darwin et al., 2003; Ives et al., 2010; Vestergaard, Fyson, & Patterson, 2009) to create the masker and thus, vary the difficulty level of the task by having several voices more or less similar to the target voice, seemed relevant. In order to be able to objectively quantify the distance between target and masker voices, masker voices were created from the target voice. The target voice is analysed and resynthesized with STRAIGHT (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999) implemented in MATLAB. The audio stimuli were originally recorded by a woman and this original voice was chosen as the target voice.

Two vocal characteristics, namely voice pitch (F0) and vocal-tract length (VTL), were manipulated during this analysis-resynthesis. F0 is modified by STRAIGHT with a shift in the overall pitch contour by a number of semitones while VTL is modified with a compression/stretching in the spectral envelope (formant peaks) of the audio signal along a linear frequency axis.

3.3.4 Semantic context

Quantifying the impact of the semantic context on voice segregation or intelligibility is difficult because it is based on the semantic distance that links words together (see Chapter 2 for more information on the semantic distance of words). The most widely adopted solution was the use of the *"Speech Perception in Noise"(SPIN)* sentences (Bilger R. C., Nuetzel J. M., Rabinowitz W. M., & Rzeczkowski C., 1984; Kalikow, Stevens, & Elliott, 1977). This test was specifically developed to include sentences where the final word can be either highly predicted from the semantic context or not. An example of a high-context sentence would be *"The boat sailed across the bay"* where words *boat*, *to sail* and *across* help participants to find the word *bay*. An example of a low-context sentence would be *"John was talking about the bay"*. The SPIN test contains only two semantic categories and does not allow the semantic context to be measured.

Recently, thanks to the contribution of the word2vec algorithm (Mikolov et al., 2013), a new approach to measure the semantic context has been created (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018; Broderick, Anderson, & Lalor, 2019), based on the numerical estimation of the semantic distance between two words (see Chapter 2). Broderick et al. (2018) calculated a word's similarity using a Pearson correlation between a word (converted to a high-dimensional vector) and the average of all the (high-dimensional vectors corresponding to the) preceding words.

Inspired by Broderick et al. (2018)'s approach and using a French word2vec model (Gaudrain & Crouzet, 2019), we created a measure of the semantic context, described below. This new semantic context measure can be defined as the probability that a word A (*i.e.* target keyword) is semantically closer to a group of words (*i.e.* target story) than a word B (*i.e.* masker keyword) and therefore tends to semantically belong to the group of words.

This measure of the semantic context assumes that the participant is listening to the target story from the beginning and can be measured as follows. Each target story is a sequence of n words t such as the target story is represented as $[T_1, T_2, T_3, \ldots, T_n]$. In a similar way, the masker story, composed of p words, is represented as $[M_1, M_2, M_3, \ldots, m_p]$. Within each target story, participants have to find three target keywords $[t_1, t_2, t_3]$ (beginning, middle or end). The three target position, indexed k, can vary from 2 to $n-1$ for the target story. Similarly, in the masker story, the indices of the three masker keywords, $m1, m2, m3$ can vary from 2 to $p\breve{}1$. The semantic similarity between two words can be estimated with word2vec (Gaudrain & Crouzet, 2019; Mikolov et al., 2013) and can be noted φ. Therefore, the semantic context can be estimated by comparing the distance between the words of the target sentence and the target keyword and the distance between the words of the target sentence and the masker keyword (see Equation 3.1).

$$\Phi_k = \frac{1}{n-1}\sum_{i \neq t_k} \varphi(T_i, T_{t_k}) - \frac{1}{n-1}\sum_{i \neq t_k} \varphi(T_i, M_{m_k}) \tag{3.1}$$

The first term of the equation 3.1 quantifies the semantic distance between a target keyword k and all the other words of the target story and varies from 0 to 1. The second term quantifies the semantic distance between the masker keyword k and all the words of the target story (except the target keyword) and varies

from 0 to 1. Therefore, Φ_k can vary from -1 to 1. In the *Long-SWoRD test 1.0*, Φ_k varies from -0.2 to 0.47.

If the semantic context Φ_k is positive, it shows that the target keyword is closer to the target story than the masker keyword. If participants are getting help from the semantic context to answer, they would then select more often the keyword from the target sentence. On the other hand, if Φ_k is negative, it may show that the masker keyword is closer to the target story than the target keyword and the participant may be biased towards the masker keyword.

3.3.5 Statistical analyses

All statistics were performed using R (R Core Team, 2017). All the generalized linear mixed models (gLMM) were implemented with the *lme4* package (Bates, Mächler, Bolker, & Walker, 2014). The models were implemented using a top-down strategy on data (Zuur, Ieno, Walker, Saveliev, & Smith, 2009). The final model is reported with the *lme4* syntax such as Equation 5.3:

$$BinaryScore \sim factor_A * factor_B + (factor_A * factor_B \mid subject) \qquad (3.2)$$

The full-factorial model is indicated by the fixed effect term $factor_A * factor_B$ and includes main effects and interactions for these two main conditions. The last term of the equation describes an individual random intercept and slope per subject for $factor_A$ and $factor_B$.

For an easier interpretation, the *afex* package (Singmann, Bolker, Westfall, & Aust, 2019) was used to compute the statistics of main effects. To do so, the final model was compared to restricted models in which the effect estimated is fixed to 0.

Finally, post-hoc analyses were computed with normalized pairwise comparisons of proportion and a *false discovery rate correction*. All the variables were normalized and centred on 0.

3.4 Experiment 1

3.4.1 Methods

Procedure

In this experiment, two competing stories were presented at the same time. The perceptual distance between the two competing stories was manipulated by modulating binaural cues and voice cues. These two types of cues are considered to be the two most important cues for speech-on-speech perception (Bronkhorst, 2015). For the spatialisation of stimuli, two presentation configurations were used in this study, namely, dichotic and diotic presentations. The signals of the target and the masker do not physically overlap in the dichotic presentation, as one is presented, through headphones, to one ear and the other to the other ear. In the diotic presentation, the signals of the target and the masker physically overlap and, in this case, listeners must rely on other perceptual cues to segregate the two speakers.

It is especially in this latter condition that vocal characteristics are important to separate two speakers. In practice, it is not easy to quantify how the voices of two individual speakers differ. Instead, based on previous studies (Başkent & Gaudrain, 2016; Darwin et al., 2003; Ives et al., 2010; Vestergaard, Fyson, & Patterson, 2009), it is more practical to operate with recordings from a single speaker, and to generate a number of voices from them by manipulating two vocal parameters: vocal pitch (F0) and the apparent length of the vocal tract (VTL).

The experiment was composed of 12 blocks from the Long-SWoRD test, each block containing 12 trials. Half of the blocks were presented diotically and the other half of the blocks dichotically. In the dichotic condition, the target speech stream was always played in the right ear. Within each presentation, three distances of voices were presented. The same distance between target and masker voices was kept within a block. The characteristics of these voices will be described in the next section. The 6 presentation conditions were randomly assigned to the blocks, which were then presented in a random order. Between each block, participants were allowed to take unlimited breaks. Data collection lasted 60 to 90 min, and the entire procedure was completed in a single session.

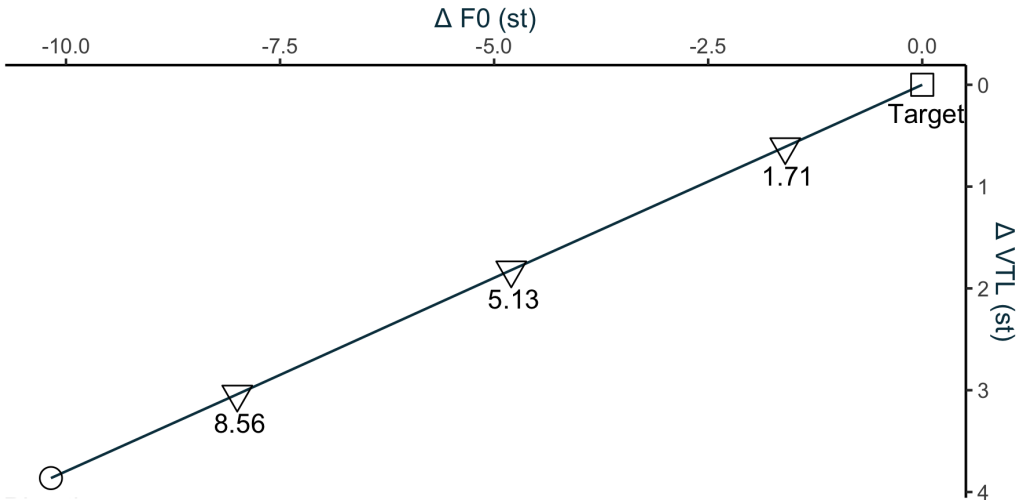

Figure 3.3 Change in semitones (st) imposed on the original target voice to create the masker voices for experiment 1. The original target voice (female acoustics) is represented by a square, the direction of the manipulated voices by a circle and the masker voices by triangles. The numbers below the triangles represent the 2-dimensional Euclidean distances between the masker voices and the original target.

Audio stimuli

The first step of the masker voices creation was to synthesize a credible male voice to obtain the direction for the changes of the original female voice (see Figure 3.3). For this purpose, the parameters F0 and VTL have been adjusted. The second step was to choose, based on literature, the parameters of two masker voices : one easy and distinguishable masker voice and one difficult and barely distinguishable masker voice. To create a "male" voice very different from a female target voice, the F0 of the original voice was shifted down of 8 semitones (st) (Başkent & Gaudrain, 2016). Accordingly with the direction fixed on the first step, the VTL of the original voice was also modified by 3.04 st. Then, to obtain a very similar female voice but still differentiable from the target voice — *a just-noticeable difference (JND) voice* — the parameters of the original voice were adjusted to obtain a total difference of 1.71 semitones along the direction axis (Gaudrain & Başkent, 2015). This total distance is calculated as $\sqrt{\Delta F0^2 + \Delta VTL^2}$. Finally, the third step was to synthesize a third voice, equidistant from the first two voices. Table 3.1 shows the parameters values for the three masker voices.

Table 3.1 Distance between the target and the masker voices in semitones for experiment 1.

Voice	ΔF0	ΔVTL	Combined distance
JND	-1.6	0.61	1.71
Intermediate	-4.8	1.82	5.13
Male	-8	3.04	8.56

Participants

There were twenty-two participants, aged between 20 and 32 ($\mu = 24$ years old, $\sigma = 3.37$). All participants were native French speakers and had audiometric thresholds ≤ 20 dB HL at audiometric test frequencies between 250 Hz and 8 kHz.

3.4.2 Results

General description

Figure 3.4 shows the average performance in the percentage of correctly identified target keywords for each condition. For each condition, all subjects demonstrated high scores, well above chance (based on a binomial test at the 5% significance level, chance is at 50%). All conditions were at the ceiling except for the diotic presentation with the just-noticeable-difference voice (1.71 st).

A generalized linear mixed model (gLMM) was fitted on the binary (correct/incorrect) scores. Equation 3.3 indicates the final model:

$$score \sim presentation * voice + (presentation + voice \mid subject) \tag{3.3}$$

Results (see Table 3.2) show that participants had better scores when stimuli were presented dichotically than diotically. There was no voice effect for the dichotic presentation whereas participants had higher scores when the distance was increasing between the target and the masker voices in a diotic presentation. For the diotic condition, post-hoc analysis confirmed that when the distance between the target and the masker voice was 1.71 st, participants had better scores than when the distance was 5.13 st [$z = -15.39, p < .001$] or 8.56 st [$z = -15.95, p < .001$]. There was no difference in performance between voices 5.13 st and 8.56 st.

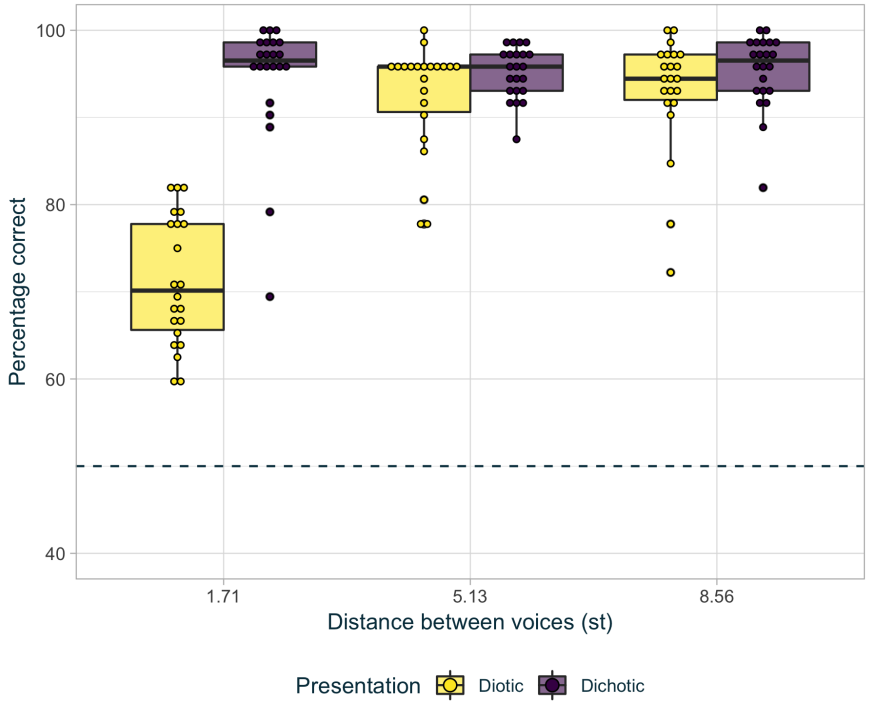

Figure 3.4 Percentage of correct responses for each voice in both diotic (yellow) and dichotic (violet) presentations. The dots represent the scores for every participant in each condition. The hinges of the boxplot represent the first and the third quartile. The median is represented as a bar in each boxplot. The length of the whiskers is 1.5 interquartile range. The dashed line (50%) indicates the level at which performance is significantly greater than chance based on a binomial test at the 5% significance level.

Keyword analysis

The position of the keyword in the story (beginning, middle and end of the story) was also analysed. The obtained results are reported in Figure 3.5. Since the voice did not have the same effect on both stimuli presentations, and, for modelling and interpretation aims, two gLMM were fitted with a top-down strategy modelling. Equation 3.4 was fitted with diotic data and Equation 3.5 with dichotic data.

$$score \sim keywords * voice + (keywords * voice \mid subject) \tag{3.4}$$

Based on the likelihood ratio tests, the position of the keyword in the sentence had an effect on participants' scores in a diotic presentation $[\chi^2(2) = 67.51, p < .001]$ as well as the voice $[\chi^2(1) = 42.47, p < .001]$ and the interaction $[\chi^2(2) = 31.06, p < .001]$. Post-hoc analysis showed that participants had higher scores when keywords were at the end of the story than at the beginning $[z = -5.34, p < .001]$ or in the

Table 3.2 gLMM coefficients for the distance between the two voices (centered on 0), the stimulus presentation and their interaction as fixed factors. β and SE are the estimated value of the coefficient and its standard error. The Wald z value and its associate p-value are the statistics of the coefficient.

Fixed effects	Coefficients		Statistics	
	β	SE	z	p
Intercept	3.2	0.17	18.86	$< .001$
Presentation	-1.06	0.15	-7.17	$< .001$
Voice	0.39	0.21	1.82	.07
Presentation × Voice	1.83	0.21	8.55	$< .001$

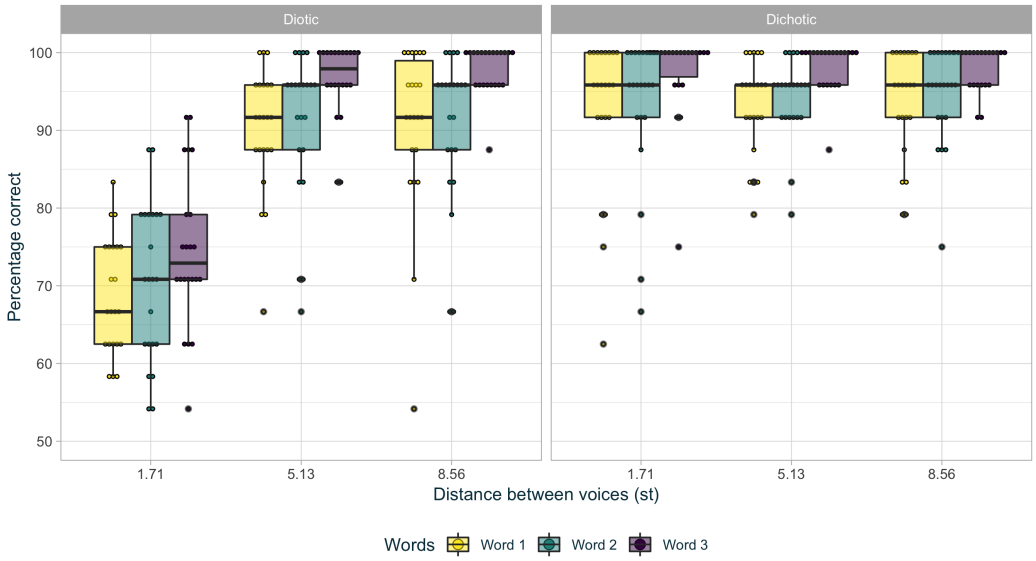

Figure 3.5 Percentage of correct responses for each voice in both diotic and dichotic presentations for every keyword.

middle of the story [$z = -4.82, p < .001$]. This effect also influenced performance significantly for voice 5.13 st and voice 8.56 st but disappeared for the voice 1.71 st, still in diotic presentation.

$$score \sim keywords + (keywords + voice \mid subject) \tag{3.5}$$

When stimuli were presented dichotically, results showed that the keyword position also influenced scores [$\chi^2(2) = 24.86, p < .001$]. Post-hoc analysis showed that participants had higher scores when keywords were at the end of the story than at the beginning [$z = -6.76, p < .001$] or in the middle of the story [$z = -5.77, p < .001$].

Error analysis

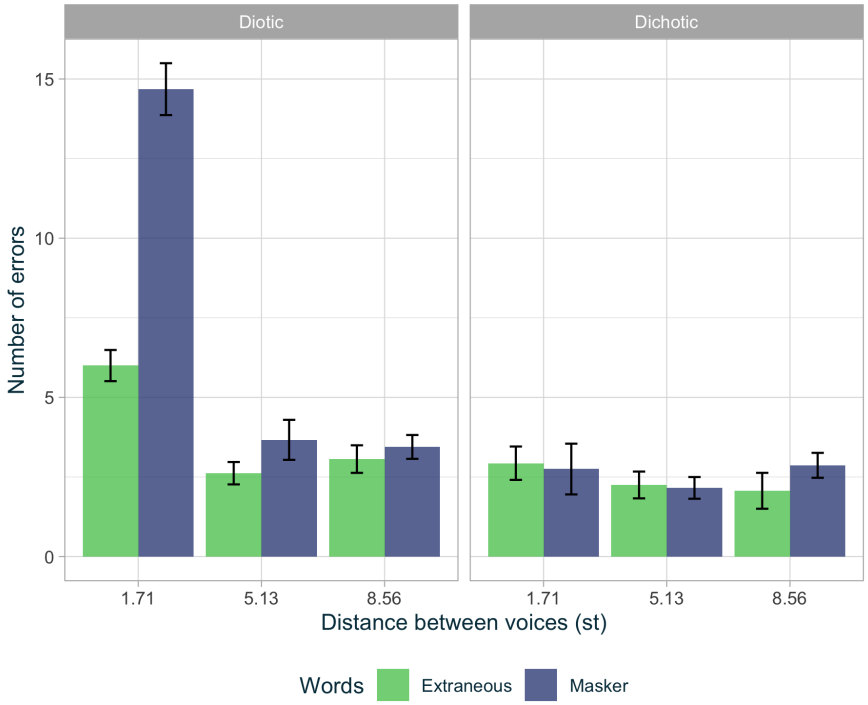

Figure 3.6 Average number of errors for each condition. The bars represent the masker answers (in blue) and the extraneous answers (in green). The error bars are the standard error of the mean.

Analysing the nature of errors is important to infer if participants were wrong because they were listening to the masker story or because they did not know which keyword to choose. Figure 3.6 illustrates the error distribution while Equation 3.6 represents the final model of a the top-down strategy modelling gLMM on binary (masker or extraneous) data from subset of answers where the participant did not respond the target keyword.

$$errortype \sim presentation * voice + (presentation * voice \mid subject) \qquad (3.6)$$

Results (see Table 3.3) show there was no voice effect for the dichotic presentation. On the contrary, in diotic presentation, participants less often chose a masker keyword when the distance between the target voice and the masker voice increases. Finally, the ratio of masker errors to extraneous errors was above chance with binomial tests at 5% significance level only when stimuli were presented diotically with the voice 1.71 st $[z = 5.34, p < .001]$. These results indicated that participants listened, at least partially, to the masker voice instead of the target voice in a

difficult condition such as a diotic presentation with a small distance between the masker and the target voice.

Table 3.3 gLMM coefficients for the distance between the two voices (centered on 0), the stimulus presentation and their interaction as fixed factors fitted on error data.

Fixed effects	Coefficients		Statistics	
	β	SE	z	p
Intercept	0.25	0.2	1.24	.21
Presentation	0.23	0.21	1.04	.3
Voice	0.46	0.35	1.32	.19
Presentation × Voice	-1.23	0.41	-3.02	$< .01$

Semantic context

A gLMM was fitted on the binary data (target or masker) with the semantic context, the distance between voices and the stimulus presentation. Equation 3.7 shows the final model with a top-down modelling:

$$score \sim voice * (presentation + \Phi_k) + (presentation + voice + \Phi_k \mid subject) \quad (3.7)$$

Regarding the stimulus presentation and the distance between voices, results were similar to previous analysis (see Table 3.4). Regarding the semantic context, only the interaction with the voice was significant. Post-hoc analysis showed that participants had better scores when the target keyword is semantically closer to the target story than the masker keyword only for voice 1.71 $[z = 3.56, p < .01]$ and voice 5.13 $[z = 2.54, p < .05]$. Figure 3.7 shows the data for the interaction between the voice and the semantic context.

3.4.3 Discussion

The advantages of spatial separation in simultaneous speech streams listening tasks found in the early literature (Broadbent, 1954; Cherry, 1953) explain the higher performance of dichotic presentation relative to diotic presentation, no matter how far away target and masker voices are.

Also in line with previous studies (e.g. Başkent & Gaudrain, 2016; Darwin et al., 2003; Ives et al., 2010; Vestergaard, Fyson, & Patterson, 2009), participant

Table 3.4 gLMM coefficients for the distance between the two voices (centered on 0), the stimulus presentation, the semantic context (centered on 0) and their interaction as fixed factors fitted on target/mask data.

Fixed effects	Coefficients		Statistics	
	β	SE	z	p
Intercept	2.69	0.14	19.27	$< .001$
Presentation	1.11	0.19	5.86	$< .001$
Voice	2.52	0.23	11.13	$< .001$
Semantic	0.83	0.44	1.9	.06
Semantic × Voice	-2.33	0.95	-2.46	$< .05$
Presentation × Voice	-2.48	0.29	-8.51	$< .001$

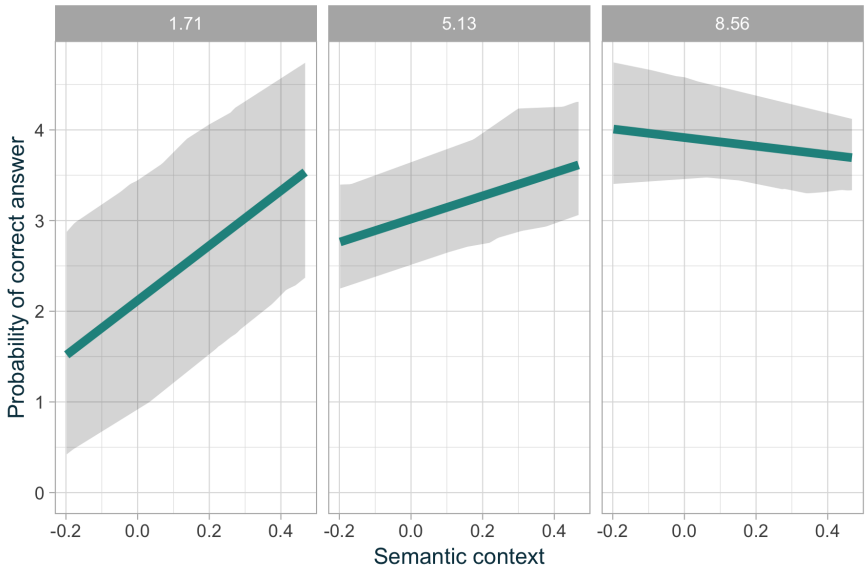

Figure 3.7 Probabilities of correct answers (with a logit transformation) per voice computed with equation 3.7. The x-axis represents the semantic context. The average probability is represented by the grey line while the first and the third quartiles are represented by the grey ribbon.

performance suffers from a decreasing distance between target and masker voices. It also seems that they can benefit from semantic context when the voices are close. However, when the distance between voices is greater than 5.13 st, participants do not seem to make use of semantic information, perhaps because using vocal characteristics is sufficient in such an easy context.

While participants' scores are capped in five of the six conditions, a diotic presentation of stimuli with a small distance between target and masker voices causes difficulties for participants. Indeed, although their performance is above

chance, mistakes were due to the choice of a keyword pronounced by the masker voice. Moreover, the advantage of the end-keyword observed under the other conditions is also not present.

The question of a performance limit then arises. What happens if the distance between the voices varies from 1.71 to 5.13 semitones? Is there a limit at which the subjects' performance suddenly drops or, on the contrary, would there be a correlation between the performance and the distance between target and masker voices? In order to address this question, a second experiment was set up.

3.5 Experiment 2

3.5.1 Methods

Procedure

In experiment 2, the procedure and the material content are similar to experiment 1. However, stimuli were presented only diotically and six voice distances were presented. Data collection lasted 60 to 100 min, and the entire procedure was completed in a single session.

Audio stimuli

The masker voices of experiment 2 were created with the same analysis- synthesis used in experiment 1. In order to be able to compare the two experiences, the voice with a just-noticeable difference (JND) has been kept (Gaudrain & Başkent, 2015). Additionally to this voice, five new equidistant voices have been synthesized. Because of the ceiling effect observed in experiment 1, it was decided that the greatest distance between target and masker voices would be 3.42 st which is equidistant to voice 1.71 st and 5.13 st from experiment 1. Parameters values for the six masker voices are displayed in Table 3.5 and Figure 3.8.

Participants

Thirty new participants (different from experiment 1), aged between 20 and 26 ($\mu = 20.94$ years old, $\sigma = 1.36$), participated in this second experiment. All of

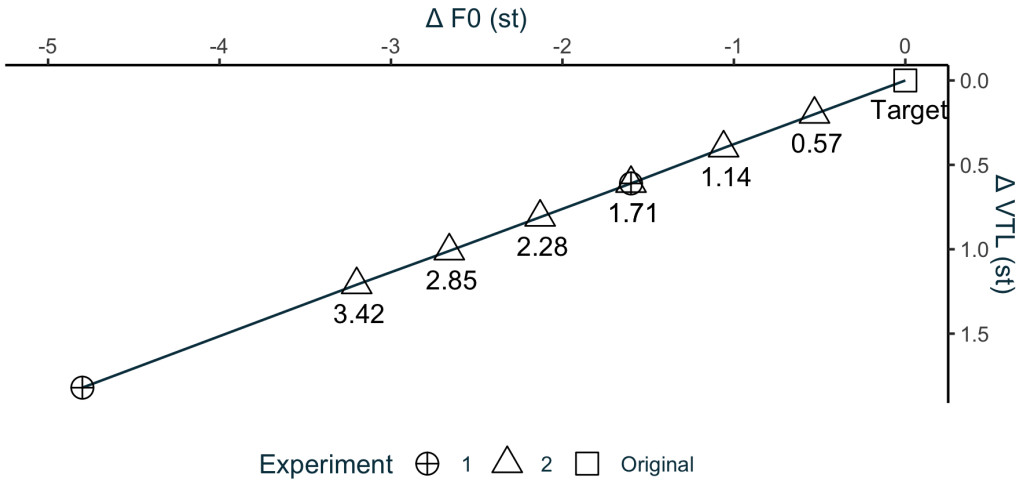

Figure 3.8 Change in semitones (st) imposed on the original target voice to create the masker voices for experiment 2. The experiment 1 masker voice is represented by circles, experiment 2 masker voices by triangles and the original female voice by a square. The voice 1.71 st is present in both experiments. The numbers below the triangles represent the 2-dimensional Euclidean distances between the masker voices and the original target.

Table 3.5 Distance between the target and the masker voices in semitones for experiment 2.

Voice	ΔF0	ΔVTL	Combined distance
1	-0.53	0.2	0.57
2	-1.06	0.4	1.14
3 (JND)	-1.6	0.61	1.71
4	-2.13	0.81	2.28
5	-2.67	1.01	2.85
6	-3.2	1.21	3.42

them were native French speakers and had audiometric thresholds ≤ 30 dB HL at audiometric test frequencies between 125 Hz and 8 kHz.

3.5.2 Results

General description

Figure 3.9 shows the average performance in the percentage of correctly identified target keywords for each condition. When masker and target voices were separated by at least 1.71 st, all subjects' results were above chance. For the voice 1.14 st, 2

participants' scores were not different from chance and for the voice 0.57, only two thirds of participants were not different from chance.

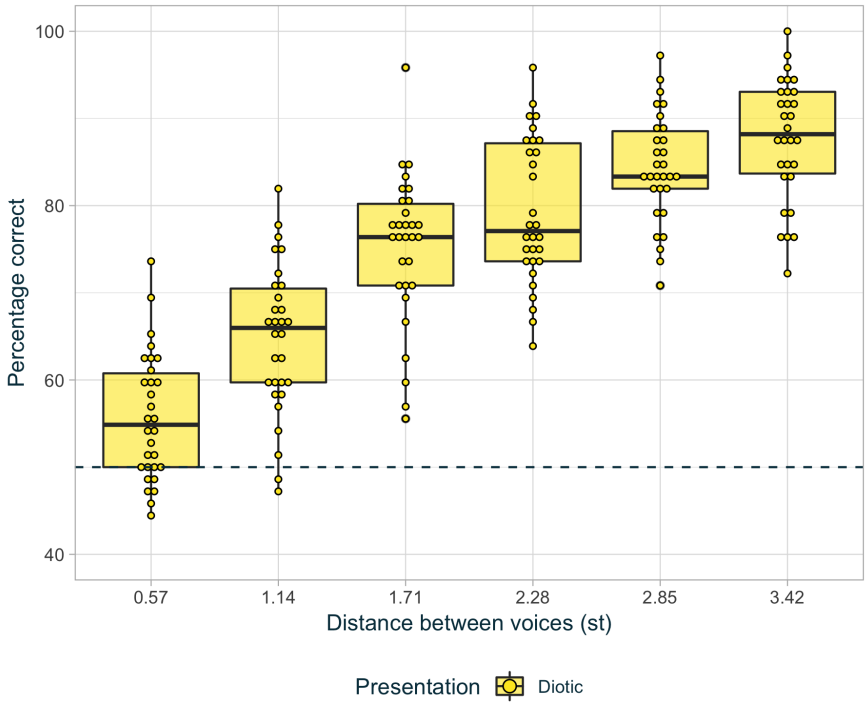

Figure 3.9 : Percentage of correct responses for each voice. The dots represent the scores for every participant in each condition. The hinges of the boxplot represent the first and the third quartile. The median is represented as a bar in each boxplot. The length of the whiskers is 1.5 interquartile range. The dashed line (50%) indicates the level at which performance is significantly greater than chance based on a binomial test at the 5% significance level.

A generalized linear mixed model (gLMM) was fitted on the binary (correct/incorrect) scores. Analysis methodology of experiment 2 is similar to experiment 1. Equation 3.8 shows the final model with a top-down strategy modelling:

$$score \sim voice + (voice \mid subject) \tag{3.8}$$

Participants had better scores when distance between the target and the masker voices is larger $[\beta = 1.9, SE = 0.12; z = 15.64, p < .001]$. Post-hoc analysis with a false discovery rate correction showed that the performance is different for each voice.

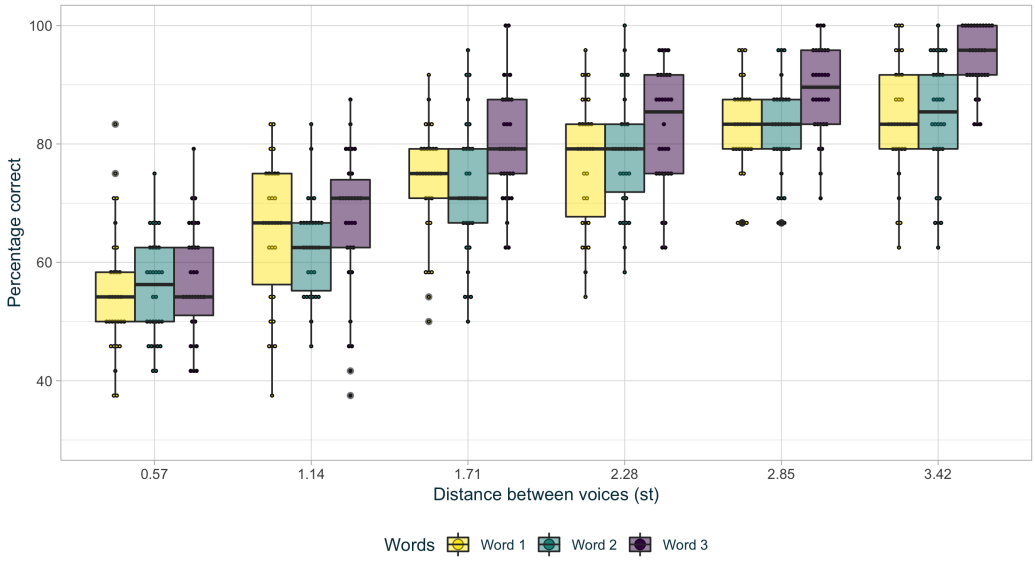

Figure 3.10 Percentage of correct responses for each voice for every keyword.

Keyword analysis

Equation 3.9 shows the final gLMM model when the position of the keyword in the story (beginning, middle and end of the story) is added the model and Figure 3.10 shows the data..

$$score \sim keywords * voice + (keywords + voice \mid subject) \tag{3.9}$$

Based on the likelihood ratio tests, the position of the keyword in the sentence had an effect on participants' scores $[\chi^2(2) = 42.41, p < .001]$ as well as the voice $[\chi^2(1) = 65.25, p < .001]$ and the interaction $[\chi^2(2) = 34.56, p < .001]$. Post-hoc analysis showed there was no performance difference for the three keywords in voice 0.57 and voice 1.14. However, there was a recency effect for the four other voices since participants have better scores for the end-keyword than for keywords in the beginning and the middle of the sentence.

Error analysis

Figure 3.11 illustrates the errors distribution and Equation 3.10 represents the final model of a top-down strategy modelling gLMM on binary (masker or extraneous) data from subset of answers where the participant did not respond the target keyword. Participants chose a keyword uttered by the masker voice over the

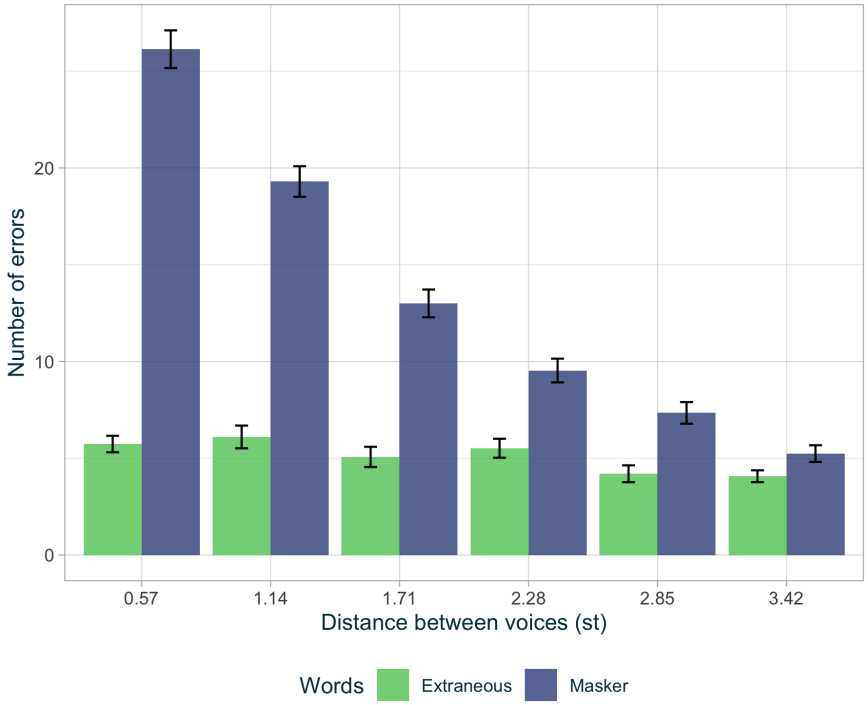

Figure 3.11 Average number of errors for each condition. The bars represent the masker answers (in blue) and the extraneous answers (in green). The error bars are the standard error of the mean.

extraneous keyword when distance between the target and the masker voices decreased $[\beta = -1.28, SE = 0.16; z = -8.12, p < .001]$.

$$errortype \sim voice + (voice \mid subject) \tag{3.10}$$

The ratio of masker errors to extraneous errors was above chance for all the voices except when the distance between voices is 3.42 semitones (see Table 3.6). Chance was computed for each voice with a binomial test.

Semantic context

A gLMM was fitted on the binary data (target or masker) with the semantic context, the distance between voices and the stimulus presentation. Equation 3.11 shows the final model with a top-down modelling:

$$score \sim voice * \Phi_k + (voice * \Phi_k \mid subject) \tag{3.11}$$

Table 3.6 Comparisons with chance level for the ratio of masker errors to extraneous errors.

Voice distance	Statistics	
	z	p
0.57	13.7	$< .001$
1.14	9.37	$< .001$
1.71	6.45	$< .001$
2.28	3.13	$< .01$
2.85	2.18	$< .05$
3.42	0.52	.6

Participants had better scores when the distance between voices is larger [$\beta = 2.34, SE = 0.15; z = 15.27, p < .001$] as well as when the measure of the semantic context is higher [$\beta = 1.24, SE = 0.23; z = 5.5, p < .001$]. The interaction between these two factors was also significant [$\beta = 1.48, SE = 0.6; z = 2.46, p < .05$]. Post-hoc analysis showed that the semantic context significantly influenced performance only for voice 1.71 [$z = 2.89, p < .05$] , voice 2.28 [$z = 2.28, p < .05$], voice 2.85 [$z = 2.82, p < .05$] and voice 3.42 [$z = 2.67, p < .05$]. Figure 3.12 shows the data for the interaction between the voice and the semantic context.

3.5.3 Discussion

Subjects' performance depends on the distance between target and masker voices. Indeed, the more the distance increases, the higher the scores are. When the distance between the voices is smaller than 2.85 st, the subjects' errors are mainly due to the selection of masker keywords and not extraneous keywords. In this case, the participants make mistakes because they listen, at least partially, to the masker. The difficulty that participants have with ignoring the masker may come from a problem of either segregating the two voices or managing to select the target voice or both at the same time. On the other hand, when the distance between the voices is over 2.85 st, the few errors are equally distributed between the masker keywords and the extraneous keywords. In this condition, participants make mistakes either because they are simply distracted or because the task is difficult. The few errors made by participants do not allow a clear interpretation of these results.

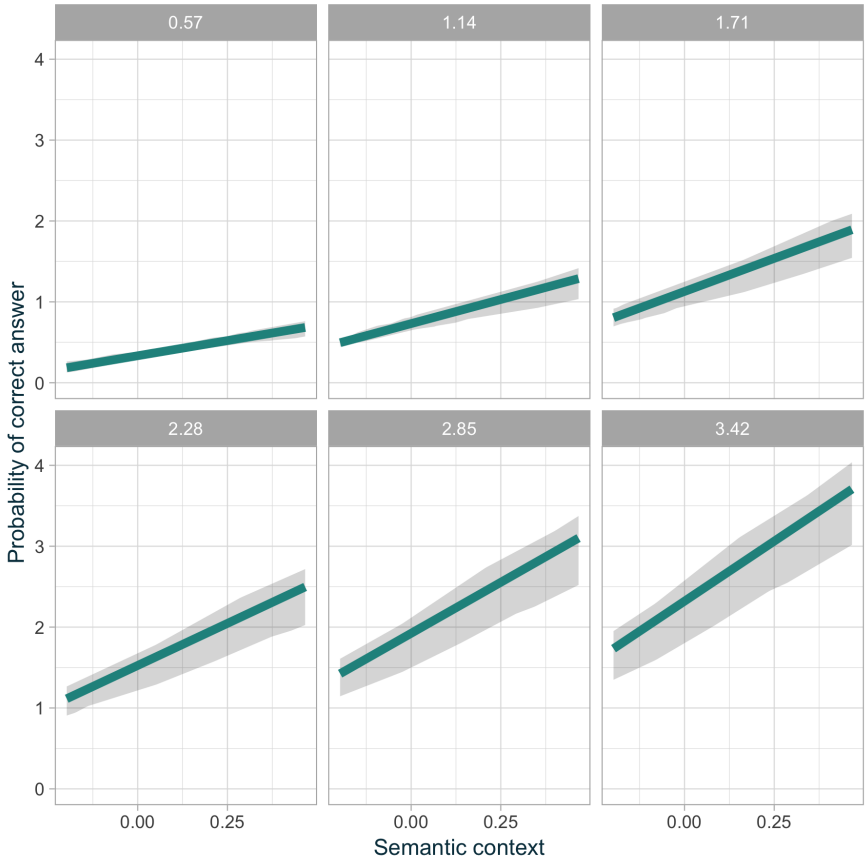

Figure 3.12 Probabilities of correct answers (with a logit transformation) per voice computed with equation 3.11. The x-axis represents the semantic context. The average probability is represented by the grey line while the first and the third quartiles are represented by the grey ribbon.

In general, participants can use a general semantic context to select the keywords being part of the target story. A possible explanation could be that all the stories come from the same audiobook and therefore share a common semantic context. However, when target and masker voices are close to each other (less or equal to 1.14 st), the participants do not seem to benefit from this information. Hence, the explanation could be that the high difficulty of these conditions makes it impossible for participants to access the semantic context. Indeed, it may be that under these conditions, participants manage to find the keywords without reaching understanding of the story, and therefore cannot fully benefit from the semantic context.

The end-keyword advantage observed in experiment 1 is also present when the distance between the voices is $\geq$ 1.71 st. Finally, the diotic presentation of the stimuli with a distance of 1.71 st between the voices is similar for experiments 1 and

2. Participants in both experiments obtained similar scores [$t(50) = -1.75, p = 0.09$]. The task proved to be consistent since the results were replicated for at least one condition.

3.6 General discussion and conclusion

The main purpose of this chapter is to examine how listeners segregate two speech streams in the context of longer stimuli, the Long-SWoRD test, in behavioural studies where both build-up effect and participants' knowledge of language are allowed to fully contribute.

In general, the experimental outcomes of the present research are consistent with the existing literature. Primitive segregation cues, such as spatialisation and vocal characteristics, clearly influenced participants' performance. The advantage of a dichotic listening condition over a diotic listening condition is consistent with previous studies (Broadbent, 1954; Cherry, 1953; Ericson & McKinley, 2001). The advantage of a large distance between voices over a small distance is also in accordance with previous studies (e.g. Başkent & Gaudrain, 2016; Darwin et al., 2003; Ives et al., 2010; Vestergaard, Fyson, & Patterson, 2009). The absolute comparison between the participants' performance with the Long-SWoRD test and participants' performance in previous studies is complicated (different procedures, different chance level, *etc.*). However, the relative comparison of results, as illustrated by the Figure 3.13, suggests that there is no difference in performance when participants perform a CRM task, a syllables discrimination task or the Long-SWoRD test (see Appendix A for more details).

Moreover, as already mentioned in previous studies (Freyman et al., 2001; Iyer et al., 2010), participants can also benefit from semantic information, only if they have managed to access the semantic context of the target story. The results obtained with the new semantic context measurement is also consistent with studies using meaningless sentences as stimuli. In addition to all the previous mentioned cues, participants can also benefit from syntactic cues as Iyer et al. (2010) already suggested. Therefore, participants might have benefited from their syntactic knowledge of language to perform the Long-SWoRD test.

Results obtained in this study also highlighted the importance of memory in a selective attention task with the presence of a recency effect for conditions where the distance between the target and the masker voices is higher than the JND.

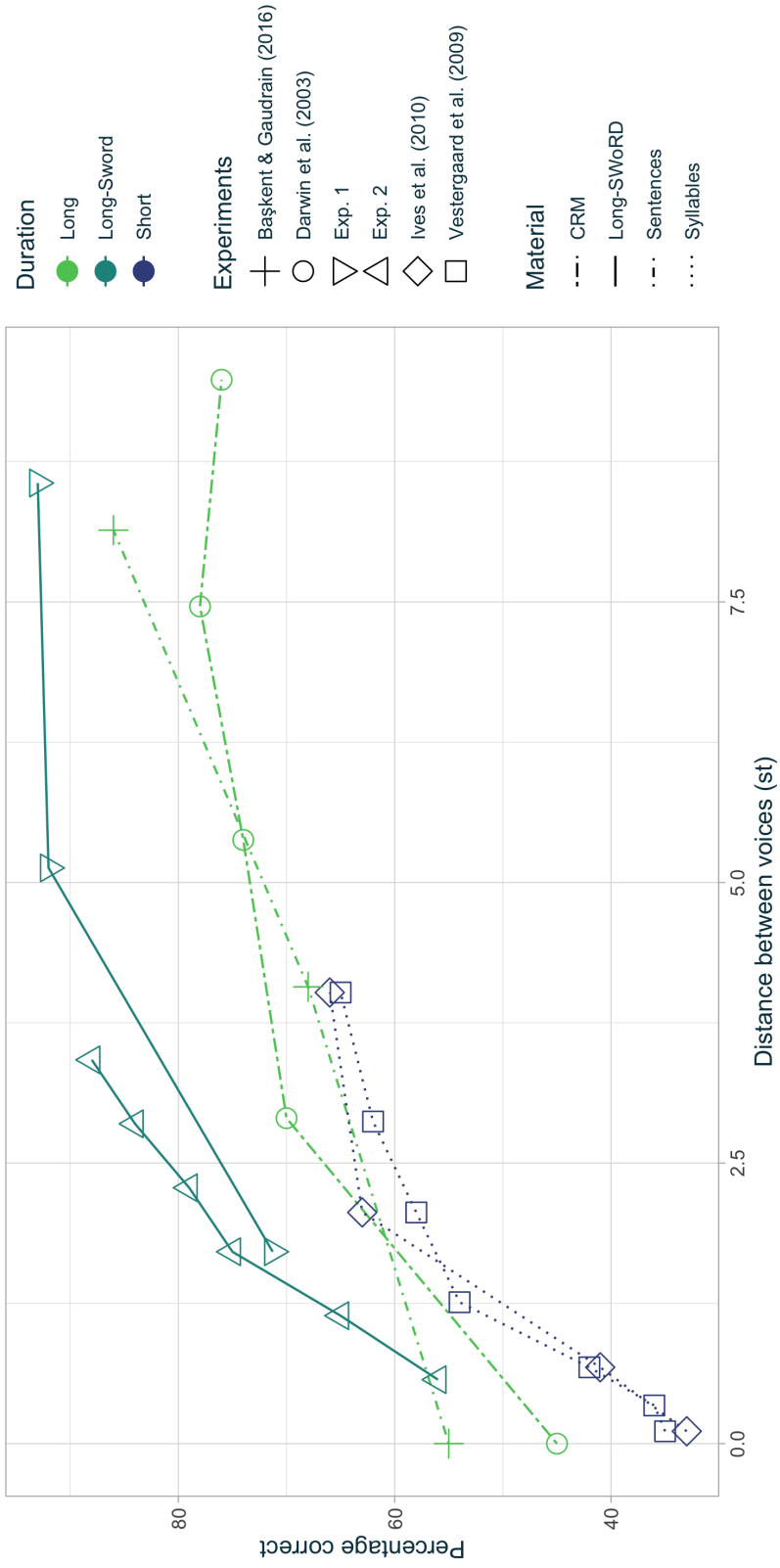

Figure 3.13 Scores for previous and current behavioural studies for vocal characteristics.

The lack of recency effect for conditions where the distance between both voices is really small could be explained by the difficulty for participants to separate both voices. In these very challenging conditions, participants selected a large amount of masker keywords. It is therefore possible that, in these difficult conditions, the participants may have a recency effect but that this effect would apply to a signal that is impossible for them to disentangle. However, it is unlikely that this is the explanation because, despite the difficulty of these conditions, the participants manage to perform the task and their results are above chance. The explanation for the lack of recency effect for these conditions may then lie in the structure of the memory. The recency effect is attributed to the added benefit auditory items receive from echoic memory (Cowan, 1984). Even though the available literature does not agree on the duration of this memory, it seems that the echoic memory decays after a few seconds (Nees, 2016). Conducting a supplementary analysis, it seems that participants benefit from that echoic memory for the last keyword since their performance is better when the last target keyword is close to the end of the sentence (see Appendix A for more informations on the analysis). According to Nees (2016), the primacy effect is related to the working memory while the auditory sensory memory, also known as echoic memory, contributes to the recency effect. Thus, in the studies presented above, participants benefit from the echoic memory but also from the working memory, by means of the inherent structure of the task. It is noteworthy that echoic and working memory occur in parallel. Therefore, the absence of primacy effect does not indicate necessarily that working memory is not required to perform the task. But the absence of primacy effect suggests that the working memory seems to not be affected by voice cues or spatialisation.

Both experiments show that the parametric control between two voices can be extended to continuous speech streams. With this approach, it is noteworthy that participants are not able to benefit from components of the voice such as the prosody used to discriminate two speakers. However, intonation (Darwin et al., 2003) and accent (Cherry, 1953) are also cues to distinguish two speakers. Since both speech streams were created from the same original voice, participants could not benefit from all the prosodic cues. For instance, participants were able to use the variations of intonation between the target and the masker sentences but not the prosodic differences that exist between two different speakers.

Finally, as mentioned above, participants have relatively similar results with the Long-SWoRD test as with other behavioural measures. Then the question of

the similarity between previous neurophysiological studies and the Long-SWoRD test is raised. Hence, the next chapter will investigate how the Long-SWoRD test can contribute to neurophysiological measurements.

Chapter 4

Neural tracking with the Long-SWoRD test

4.1 Introduction

The ability of the human brain to separate concurrent voices is fascinating and the mechanisms involved in this endeavour remain largely unknown. Many studies have focused on situations where the two competing voices are clearly separated (male vs. female, presented dichotically). However, to study the mechanisms involved in speech separation, we are interested in situations where segregating the target speech from the masker is much more challenging and will at least partially fail. For that purpose, we have designed an experiment based on the Long-SWoRD test where the voice distance between target and masker is manipulated parametrically. With this approach, we were able to highlight specific mechanisms that are involved only when the two voices are clearly distinct, or only when the two voices are very similar to each other. As already mentioned in Chapter 1, studies investigating speech processing can use many imaging techniques (*e.g.* fMRI, EcoG, *etc.*). In this introduction, studies involving only two competing talkers and using methods that have excellent temporal resolution (electroencephalography (*EEG*) and magnetoencephalography (*MEG*)) are presented.

4.1.1 High temporal resolution methods for speech processing imaging

Event-related potentials (ERPs)

Traditionally, EEG and MEG have been used to record and measure electrical or magnetic cortical responses that are time-locked to the presentation of stimuli. Usually, these measurements are averaged over a large number of trials in order to eliminate noise from other sources (muscle artefacts, eye movements, etc.). The obtained patterns are referred to as *event-related potentials (ERP)* and are usually interpreted to reflect different stages in the sensory and cognitive processing of speech streams. Usually, earlier responses, up to roughly 100 milliseconds after stimulus, are classified as "sensory" as they depend largely on the physical parameters of the stimulus (Sur & Sinha, 2009). In contrast, the later responses are identified as "cognitive" and reflect the manner in which the subject evaluates the stimulus.

Besides the latency, the amplitude of the waveform is also important to describe the ERP. For instance, if negativity is observed about 100 ms after the onset of the stimulus, the N100 (also known as N1) indicates that the stimulus presented matches previously experienced stimuli. If positivity is observed 300 ms after the onset of the stimulus, the P300 (or P3) is interpreted as a sign that the stimulus has received a certain attention. In the same way, the N100 is thought to be a marker of semantic and lexical processing while the P600 is related to the syntax (for more information, see Sur & Sinha, 2009).

The analysis of, ERPs however, is difficult for longer stimuli that may have multiple onsets. For instance, the words "blue" and "orange" differ in duration by at least 100 ms (which is the latency of the earlier ERP). The more complex stimuli, such as sentences, the more arduous analysing and interpreting ERPs. How to disentangle an N4 for one word from an N1 for the next word? New analysis techniques have been developed in order to address this question.

Temporal Response Function (TRF)

Inspired by the latest advances in the field of vision, Lalor, Power, Reilly, and Foxe (2009) published a new analysis method that eliminates the restriction of short, simple and discrete stimuli: the *"Auditory-evoked spread spectrum analysis*

(AESPA)". The estimation of the AESPA consists in finding the unknown impulse response that links the cortical data and the audio stimuli by performing a linear least-squares estimation. Using broadband noises, Lalor et al. (2009) showed that AESPAs, although not equivalent to ERPs, share with the latter a number of properties including detailed temporal precision.

Afterwards, Lalor's team extends this method to continuous speech streams (Lalor & Foxe, 2010) and then to cocktail party situations (Power, Foxe, Forde, Reilly, & Lalor, 2012). In the latter study, the authors reported a difference between the AESPAs for the attended and unattended speech streams around the 200 ms latency in the left hemisphere. The authors hypothesized that this difference could be interpreted as an attentional suppression of the unattended stream before cognitive functions such as semantic processing occur.

Two years later, the AESPA was renamed *"Temporal Response Function (TRF)"* when it was associated with the STRF of more invasive neuroimaging techniques (see Chapter 1 for more details). At that time, O'Sullivan et al. (2015) showed that TRFs can be exploited to reconstruct the attended streams, and to a lesser extent, the unattended streams. The evaluation of the comparisons of these reconstructed streams with the original stimuli was at the core of the *"Auditory attention decoding" (AAD)* method. For instance, O'Sullivan et al. (2015) observed a median Pearson's correlation of 0.054 between the original and the reconstructed attended speech streams while the median Pearson's correlation between the original and the reconstructed unattended speech streams only yielded -0.005. Although these correlations are weak and therefore reflect a poor reconstruction of the speech signal, they are sufficient to lead to an AAD of 89% since the latter is calculated by comparing for each trial the two Pearson's correlations (see Chapter 5 for more information).

Generally, this line of research shows there is a measurable correlation between the auditory stimuli reconstructed from TRFs and the cortical oscillations. More specifically, brain oscillations synchronize with the amplitude envelope of the speech (also known as the temporal envelope). Models that link brain and stimulus are known as the *"neural"* or *"cortical tracking"*, sometimes even *"neural entrainment"* (see Chapter 1 for more precision). The accuracy of these models is evaluated through a correlation between the original and reconstructed stimuli known as *Stimulus-reconstruction* evaluation or sometimes as *neural tracking accuracy*. Until now, optimizing TRF and more generally improving the reconstruction of audio stimuli has been in the centre of the TRF neural tracking literature. The next

Chapter 5 will focus on the enhancement of these neural tracking measures while this Chapter is exploiting the TRF methodology to study brain mechanisms involved in speech-on-speech perception.

4.1.2 Neural tracking and speech streams segregation

From mixture to separate streams

Before the two speech streams are separated, the brain perceives them as a single mixture (see Chapter 1 for more information). In a MEG study, Puvvada and Simon (2017) investigated the hierarchical decomposition of this mixture into two streams. They showed that early neural responses (before 85 ms) represent the entire acoustic scene holistically (the mixture), rather than individual speech streams. At this latency, attended and unattended speech streams are equally represented in the primary acoustic area. On the contrary, in later neural responses (later than 85 ms), the attended speech stream is better represented than the unattended speech stream. These results were also reported by Brodbeck, Hong, and Simon (2018) who observed a peak in the mixture TRF at 70 ms and a peak in the attended TRF at 150ms. Another study conducted by Hausfeld, Riecke, Valente, and Formisano (2018) examined the stimulus-reconstruction of a mixture composed of three streams (two speech streams and one music stream). They also reported results consistent with those of Puvvada and Simon (2017) and in addition, showed that the grouping of two unattended streams can occur outside the listener's focus of attention, based on the acoustic similarities of the unattended streams.

The process of segregating the mixture into different streams can be helped by several acoustic and linguistic cues (for more information, see Chapter 1 and Chapter 3). The next two sections therefore report on the studies that focused on these different elements.

Acoustic cues

The sound level difference between two speech streams or *target-to-masker ratio (TMR)* is one of the main cues that helps segregate two talkers. In a recent study, Fiedler, Wöstmann, Herbst, and Obleser (2019) analysed the brain's response in a cocktail party situation while the TMR varied dynamically between -6 and 6 dB.

Overall, they found that the stimulus-reconstruction was dominated mainly by the attended speaker while under most demanding listening conditions, the AAD was dominated by the unattended talker. They also observed the presence of a $P1_{TRF}/N1_{TRF}/P2_{TRF}$ pattern for the attended stream while only a $P1_{TRF}$ was detected for the unattended stream. In addition, a $N2_{TRF}$ was present for the unattended stream under the most adverse listening condition, and was thought to be a marker of an active suppression of the unattended stream. The authors hypothesized that the early neural filters tuned to the spectro-temporal properties of the attended talker were not sufficient enough to resolve the acoustic scene. A later filter, around 250 ms, on the unattended signal is needed to actively suppress distracting inputs.

The fundamental frequency (F0) of the two voices is also an important cue for separating two speakers. A methodological paper (Teoh, Cappelloni, & Lalor, 2019) showed that pitch processing is represented in the band limited to the delta frequency range of the EEG (from 1 to 4 Hz).

Linguistic information

The major advantage of TRFs over ERPs is the use of long stimuli that allows a finer understanding of the linguistic context contribution in talker segregation (also called schema-based segregation by Bregman). Therefore, the linguistic cues were the most examined using TRFs.

As already mentioned previously in Chapter 3, Broderick et al. (2018) quantified the meaning carried by words with a semantic context measure based on semantic similarity. They incorporated this measure into TRFs and showed that there was a prominent negativity at a lag of about 400 ms when subjects understood the speech they heard. This N400_{TRF} was similar to N400_{ERP}. Figure 4.1 shows that the N400 is apparent for attended speech but not for unattended speech. Broderick et al. (2019) showed in another study that a word was better encoded when it was more similar to its semantic context.

Di Liberto, O'Sullivan, and Lalor (2015) showed that the activity in non-primary auditory cortex was modulated by phonetic features from 100 to 200 ms. In a MEG study, Brodbeck et al. (2018) analysed the contribution of several speech components (acoustic, phonetic and lexical). They also reported a brain response latency of 114 ms for phonetic information. Their results also suggest that in a cocktail party paradigm, only the attended speech is processed at the

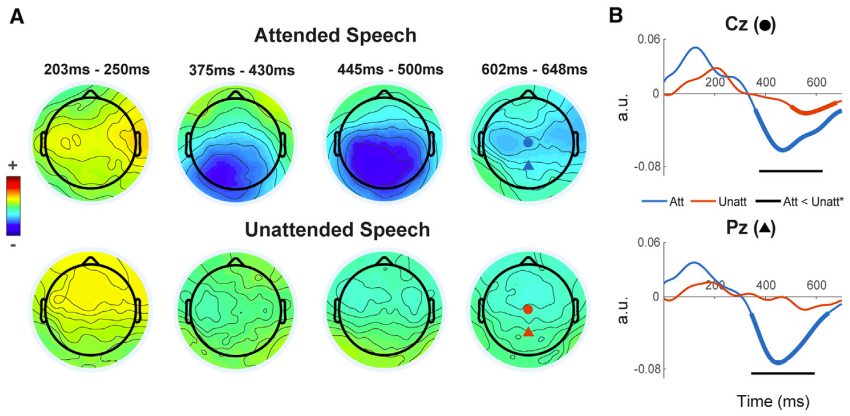

Figure 4.1 : A) Topographic maps of the TRF for attended and unattended speech. There is a centro-parietal negativity around 375 and 500 ms (N400) for the attended speech. B) Grand average TRF waveforms for both speech streams over two selected midline electrodes (adapted from Broderick et al., 2018).

lexical level. Unattended words might be identified, but their properties are not processed.

In conclusion, few studies have so far investigated the linguistic cues that facilitate the analysis of auditory scenes composed of two speakers. However, the first insights suggest that the unattended speech stream, unlike the attended speech stream, is not processed at a semantic or lexical level by the brain. The next section will focus on top-down mechanisms that help segregate the speech streams.

4.1.3 TRF and top-down mechanisms

Di Liberto, Crosse, and Lalor (2018) investigated the impact of prior knowledge on the cortical tracking of continuous speech stream. They presented degraded speech sentences to participants that were either preceded by a clear recording of the same sentences or not. When the prior information was available, the stimulus-reconstruction of the speech was reduced. These unexpected results indicated that prior information does affect the cortical encoding of speech stimuli even though it is unclear how this effect comes about. One interpretation of the authors was that the reduction of the neural tracking results from the increase of the stimulus predictability.

Other factors, such as working memory, have also been shown to affect speech tracking ability. Hjortkjaer, Märcher-Rørsted, Fuglsang, and Dau (2018) measured

the neural tracking of continuous speech while participants were performing an auditory n-back task with noise. Their results showed that when the working memory load was increasing, the stimulus-reconstruction of the speech was reduced. According to the authors, this reduction in speech entrainment is explained by a reallocation of cognitive resources when the task is demanding. More specifically, when the difficulty of the task is increasing, the attentional focus toward the phonological loop is also increased, leading to fewer resources for the cortical tracking of the ongoing stimulus. In addition, even though increasing the n-back level of the task reduces the stimulus-reconstruction evaluation, there was no difference in the TRF. On the other hand, when the background noise level was increased, the TRF peak around 150 ms was attenuated and delayed.

These studies indicate that neural entrainment is not a passive following of the speech signal. Top-down inputs such as prior information or working memory can influence it.

4.2 Rationale

As mentioned in section 4.1, the TRFs overcome the stimuli duration limitation of neurophysiological studies. Despite the different types of stimuli between ERP studies (simple and short like words) and TRF studies (typically 1-minute long audiobook extracts), common patterns seem to emerge. This is for example the case with the observation of neuronal markers such as N100 or N400.

However, as noted in previous chapters, listening to these two types of stimuli can hardly be comparable in at least two aspects. Firstly, involved cognitive resources are not identical in the two types of experiments (see Chapter 1). It seems noticeably more complicated to keep one's attention focused constantly for several tens of seconds on a target stream while a masker stream is being played concurrently than when the cognitive effort is focused on stimuli of a few seconds. Second, the nature of the task may also be questionable. With a few exceptions (*e.g.* Crosse, Butler, & Lalor, 2015), the participants' task in TRF studies is to answer a multi-choice questionnaire on the content of the story after one minute. It appears that in the case of TRFs, listening to a story is more akin to a comprehension task, whereas listening to a single word may be more a matter of intelligibility (see Chapter 2).

A few experiments (Decruy, Vanthornhout, & Francart, 2019; Lesenfants, Vanthornhout, Verschueren, Decruy, & Francart, 2019; Vanthornhout, Decruy, Wouters, Simon, & Francart, 2018) tried to bridge the gap between intelligibility and TRFs by attempting to predict speech intelligibility from the neural tracking in speech-in-noise studies. Overall, these studies reported a relation between the speech reception threshold (*SRT* or the SNR at which the subject can understand 50 % of the words) and the strength of cortical tracking: the easier it was to listen, the better the reconstruction of the stimulus. However, the stimuli used to compute the TRF and the neural tracking were not presented in the same condition: whereas the TRF was trained with a long story in silence, the tracking was assessed with Matrix sentences in noise conditions. These two listening situations are different in at least two ways (cognitive resources and of the nature task) and inferring neural tracking on Matrices from TRFs calculated in a passive listening situation could lead to inaccurate measurements.

Several questions arise from these observations. First, to what extent does the nature of the stimuli influence neurophysiological measures? Will the results be similar to the TRF literature if participants take the *Long-SWoRD* (see Chapter 2) test rather than a comprehension task on one-minute stories? Or on the contrary, will we observe patterns more similar to ERPs studies? *A priori*, the neural tracking should be different depending on the task because Hjortkjaer et al. (2018)'s study suggested that cortical tracking can be influenced by working memory.

Second, is it possible to estimate intelligibility through neural tracking measures? To the extent that TRFs are trained in the same listening condition, will there be a relationship between the behavioural performance of the participants and the neural measures? In general, a better reconstruction of the target than of the masker is observed. We can therefore assume that in easy listening situations, the same pattern will be obtained. On the other hand, in more adverse listening situations, the masker reconstruction should be greater while the target reconstruction should be smaller. Therefore, there should be a correlation between subject performance in the *Long-SWoRD* test and neural tracking measures.

To address these questions, we recorded and analyse with TRFs participants' electroencephalographic activity while they were taking the Long-SWoRD test with varying degrees of adverse listening conditions. In this chapter, the listening difficulty of the Long-SWoRD test is modulated by modifying the vocal characteristics of the masker (see Chapter 3). This is an important and extensively studied

cue which can be used to disentangle target from masker speech streams (*e.g.* Brungart et al., 2001; Darwin et al., 2003; Vestergaard, Fyson, & Patterson, 2009). Three levels of difficulty were set in this experiment. An easy level in order to replicate the results of the literature (a larger reconstruction of the target than of the masker) and a very difficult level where the neural tracking of the target and the masker should be more equivalent to each other. However when target and masker voices are similar and, hence the task is difficult, attributing participant errors to poor voice segregation or auditory attention problems can be challenging as these two phenomena can be difficult to disentangle (see Chapter 1). A third, intermediate condition was therefore required to tackle this issue. This condition remains relatively easy for the participants (hence, segregation of the voices should occur) but requires more cognitive resources. Neural tracking is considered to be measure of selective attention, so target and masker reconstructions should be somewhere between easy and difficult conditions.

Apart from acoustic cues, the segregation of two voices can also be studied under the scope of linguistic cues. For instance, Broderick et al. (2018) reported the presence of a neural semantic marker in the target tracking. Thus, the linguistic contributions of the participants is examined *a posteriori* with a measure of the semantic context, developed and presented in Chapter 3.

Finally, to relate to the *Auditory Scene Analysis (ASA)* (see Chapter 1), we investigated the streaming phenomenon (when the mixture turns into two distinct streams). Puvvada and Simon (2017) already investigated the hierarchical segregation of the mixture into two streams with TRFs and showed that early neural responses represent the mixture rather than two separate speech streams. The phase of decomposition increases over time and is usually referred as the *build-up effect* (Bregman, 1978). Thus, we examined the evolution of the build-up effect over time with TRFs.

4.3 Methods

4.3.1 Participants

Twenty-one participants, aged between 19 and 25 ($\mu = 21$ years old, $\sigma = 1.76$), participated in the experiment. All of them were native French speakers and had audiometric thresholds ≤ 30 dB HL at audiometric test frequencies between

125 Hz and 8 kHz. Participants gave informed consent before taking part in the study.

4.3.2 Stimuli and procedure

The procedure and the stimuli creation were identical to previous experiments described in Chapter 3. Subjects undertook the *Long-SWoRD test* coupled with the audio stimuli from the audiobook *Le Charme discret de l'intestin* [The Inside Story of Our Body's Most Underrated Organ] (Enders et al., 2016) (see Chapter 2). Data collection lasted 60 to 100 min, and the entire procedure was completed in a single session. Participants were instructed to avoid eye movements to reduce potential noise in the EEG recording. Stimuli were presented with OpenSesame (Mathôt et al., 2012). Participants listened to stimuli diotically over a Sennheiser HD250 Linear II headphone in a sound-attenuated booth.

Two competing stories were presented diotically at the same time. The perceptual distance between the two competing stories, hence the difficulty level of the task, was manipulated by the parametric distance in semitones (st) between the target and the masker stories. The target voice was analysed and resynthesized with STRAIGHT (Kawahara et al., 1999) implemented in MATLAB. The voice pitch (F0) and vocal-tract length (VTL) were manipulated to create the masker voices. Three conditions were needed to investigate our hypotheses: a difficult, an intermediate and an easy condition. For the hardest condition, a distance of 1.14 st between the two voices was chosen. In experiment 2 of Chapter 3, it was the condition where participants selected the most masker words while having a higher success rate than chance. In other words, participants managed to complete the task while having their attention drawn to the masker voice. For the easiest condition, a distance of 5.13 st was chosen between the two voices since the conclusions of Chapter 3 pointed to a ceiling in performance beyond 5 semitones. In this easy condition, participants' performance was over 90% and the errors were evenly distributed between masker and extraneous keywords. Finally, a distance of 3.42 st between the voices was chosen for the intermediate condition. In this intermediate condition, participants did well on the task but not as well as in the easy condition [$t = 2,66, p < .01$]. Similarly to the easy condition, participants made as many masker errors as extraneous errors. Parameter values for the three masker voices are displayed in Table 4.1 and Figure 4.2.

Table 4.1 Distance between the target and the masker voices in semitones for experiment 3. The original voice (target) is represented by a square, the direction by a circle and the masker voices by triangles

Voice	ΔF0	ΔVTL	Combined distance
Difficult	-1.06	0.4	1.14
Intermediate	-3.2	1.21	3.42
Easy	-4.8	1.82	5.13

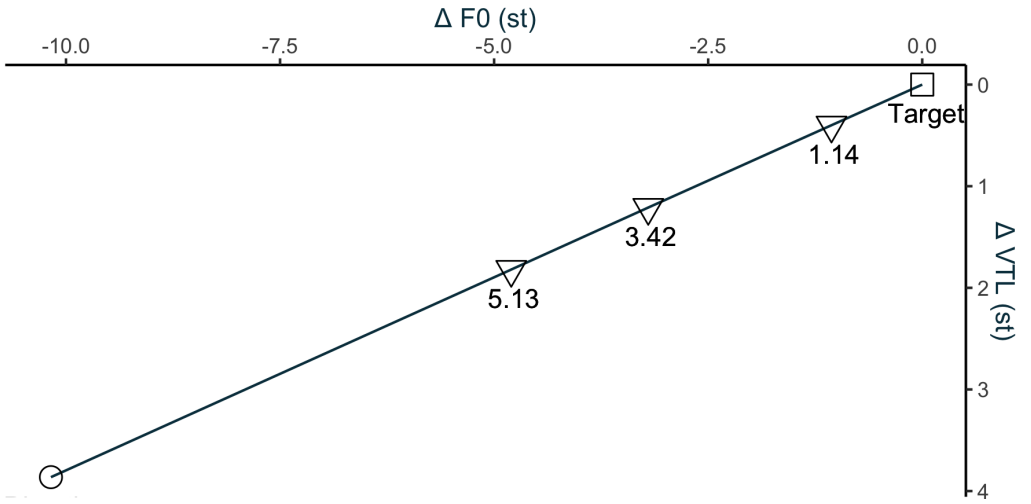

Figure 4.2 Change in semitones (st) imposed on the original target voice to create the masker voices for experiment 3. The original target voice (female acoustics) is represented by a square, the direction of the manipulated voices by a circle and the masker voices by triangles. The numbers below the triangles represent the 2-dimensional Euclidean distances between the masker voices and the original target.

4.3.3 Data acquisition and preprocessing

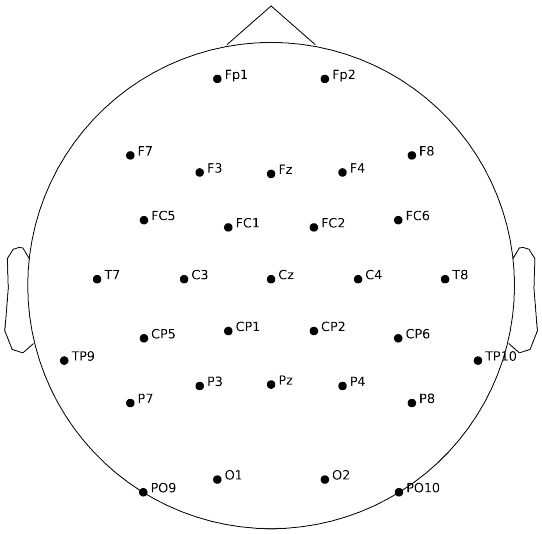

Figure 4.3 Channel positions. All the channels are distributed across the frontal (F), the parietal (P), the temporal (T) and the occipital (O) lobes. The letter C stands for "central" area.

Electroencephalographic data were recorded using ActiCap (Brain Products, Munich, Germany) with a setup of 31 channels (see Figure 4.2) at a sampling rate of 1000 Hz. Then, following O'Sullivan et al. (2015), EEG data were band-pass filtered between 2 and 8 Hz. Finally, to decrease processing time, EEG recordings were downsampled to 64 Hz.

The stimulus speech envelope was extracted with a gammatone filterbank (Søndergaard & Majdak, 2013; Søndergaard, Torrésani, & Balazs, 2012) followed by a power law (see Chapter 5 for more details). This gammatone filterbank mimics the auditory filters present in the cochlea. The gammatone filter was used with 28 subbands centred on frequencies from 50 until 5000 Hz, equally spaced on the ERB scale. The envelope in each band was extracted by taking the absolute value and then raising it to the power 0.6. A combined envelope was then computed by averaging all the 28 envelopes. The speech envelope was then downsampled to 64 Hz and low-pass filtered below 8 Hz, following the method described by O'Sullivan et al. (2015).

4.3.4 TRF and Stimulus-reconstruction

The TRF was calculated using the MNE-Python library (Gramfort et al., 2014). The TRF, also called *decoder*, is composed of weights that can be estimated by linear regression for a set of N electrodes at different delays τ. In this experiment, we investigated time delays between -900 to 0 ms. The reconstruction of the speech envelope $\hat{S}(t)$ can be obtained as follows:

$$\hat{S}(t) = \sum_{n=1}^{N} \sum_{\tau} TRF(\tau, n) R(t - \tau, n) \tag{4.1}$$

where R represents the matrix that contains the shifted neural responses of each electrode n at time $t = 1...T$. A ridge regression (see Chapter 5) can be applied to obtain the weights of the TRF as follows:

$$TRF = (RR^T + \lambda I)^{-1}(RS^T) \tag{4.2}$$

where λ is the regularization parameter, chosen to optimize the stimulus- response reconstruction, and I is the identity matrix. The optimal ridge parameter was set to $\lambda = 10^{1/2}$ accordingly to Crosse, Di Liberto, Bednar, and Lalor (2016).

TRFs were estimated on a trial-by-trial basis for each subject in each condition for the target and masker streams. The stimulus-reconstruction of a single trial was predicted in a leave-one-out fashion. That is since each subject performed 48 trials per condition, each trial was reconstructed with the TRF obtained from training the model with data from the on the 47 other trials.

The stimulus-reconstruction was evaluated with the Pearson's correlation coefficient, R, between the reconstructed speech envelope and the original speech envelope. The auditory attention decoding accuracy is the percentage of trials where a talker is correctly identified as being attended or ignored. For instance, when the TRF is trained with the target envelopes, a trial is deemed correctly classified if the reconstructed envelope is more correlated with the target than with the masker envelope. On the other hand, if the TRF is trained with the masker envelopes, a trial is deemed correctly classified if the reconstructed envelope is more correlated with the masker than with the target envelope.

To ensure that the observed effects are not just random occurrences, each reconstructed envelope was also compared to the 96 envelopes from the two other

conditions, on which the TRFs had not been trained. The surrogate data are the averaged correlation across the 96 comparisons and were used as chance level.

4.3.5 Semantic Context

The semantic context was quantified using a method adopted in previous experiments described in Chapter 3. The semantic context Φ_k can be estimated by comparing the distance between the words of the target sentence $[t_1, t_2, t_3, \ldots, t_{n_1}]$ and the target keyword t_k and the distance between the words of the target sentence and the masker keyword m_k using word2vec (Gaudrain & Crouzet, 2019; Mikolov et al., 2013) as follows:

$$\Phi_k = \frac{1}{n_1} \sum_{i \neq k} \varphi(t_i, t_k) - \frac{1}{n_2} \sum_{i \neq k} \varphi(t_i, m_k) \tag{4.3}$$

where k represents the keyword k (beginning, middle or end) and φ the similarity between two words. This semantic context measure can be defined as the probability that a word A (the target keyword) is semantically closer to a group of words (the target story) than a word B (the masker keyword) and therefore tends to belong semantically to the group of words. If the semantic context Φ_k is positive, it shows that the target keyword is closer to the target story than the masker keyword. If participants are getting help from the semantic context to answer, they would then select more often the keyword from the target sentence. On the other hand, if Φ_k is negative, it may show that the masker keyword is closer to the target story than the target keyword and the participant may be biased into answering the masker keyword.

4.3.6 Statistical analyses

All statistics were performed using R (R Core Team, 2017).

The (generalized) linear mixed models ((g)LMM) were implemented with the *lme4* package (Bates et al., 2014). The models were implemented using a top-down strategy on data (Zuur et al., 2009). The final model is reported with the *lme4* syntax such as Equation 4.4:

$$(Binary)Score \sim factor_A * factor_B + (factor_A * factor_B \mid subject) \tag{4.4}$$

The full-factorial model is indicated by the fixed effect term $factor_A * factor_B$ and includes main effects and interactions for these two main conditions. The last term of the equation describes an individual random intercept and slope per subject for $factor_A$ and $factor_B$. For an easier interpretation, the *afex* package (Singmann et al., 2019) was used to compute the statistics of main effects. To do so, the final model was compared to restricted models in which the effect estimated is fixed to 0. post-hoc analyses were computed with normalized pairwise comparisons of proportion and a *false discovery rate correction*. All the variables were normalized and centred on 0.

The generalized additive models (GAM) were used to assess non-linear function (i.e. smooth) and were performed with the *mgcv* package (Wood, 2017) and interpreted with the package *itsadug* (van Rij, Wieling, Baayen, & van Rijn, 2017). The model is reported with the *mgcv* syntax such as Equation 4.5

$$Score \sim s(var_A, by = var_B) + var_B + s(var_A, subject, bs = fs, m = 1, by = var_B) \tag{4.5}$$

The function *"s()"* is used to fit a one-dimensional non-linear regression line (or spline) across var_A. If the argument $by{=}var_B$ is present, it indicates that the non-linear regression line has to be estimated for each level of var_B. At the same time, the term var_B indicates there is a different intercept for each level of var_B. Finally, a random smooth factor (*"bs=fs"*), indicated by the term $s(var_A, subject, bs = fs, m = 1, by = var_B)$, adjusts the regression line per *subject*. Conventionally, (*"m=1"*) is also specified for shrinkage and reducing the co-linearity between the global and the subject-specific smoothers.

4.4 Results

4.4.1 Behavioural results

General description

Figure 4.4 shows the average performance as the percentage of correctly identified target keywords for each condition. All subjects' results are above chance.

A generalized linear mixed model (gLMM) was fitted on the binary (correct/incorrect) scores. Equation 4.6 shows the final model with a top-down

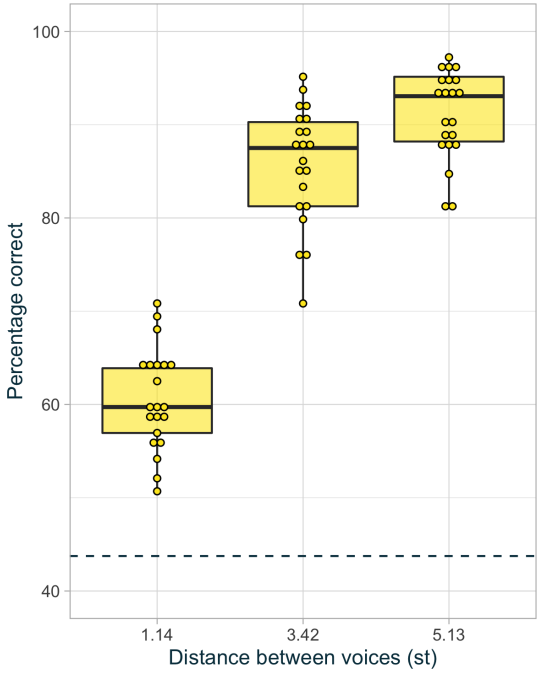

Figure 4.4 : Percentage of correct responses for each voice. The dots represent the scores for every participant in each condition. The hinges of the boxplot represent the first and the third quartile. The median is represented as a bar in each boxplot. The length of the whiskers is 1.5 interquartile range. The dashed line (43.75%) indicates the level at which performance is significantly greater than chance based on a binomial test at the 5% significance level.

strategy modelling:

$$score \sim voice + (voice \mid subject) \tag{4.6}$$

Participants had better scores when the distance between the target and the masker voices is larger $[\beta = 2.13, SE = 0.12; z = 18.02, p < .001]$. Post-hoc analysis with a false discovery rate correction showed that average scores in each voice condition were all different from one another.

Keyword analysis

Equation 4.7 shows the final gLMM model when the position of the keyword in the story (beginning, middle and end of the story) is added the model and Figure 4.5 shows the data.

$$score \sim keywords * voice + (1 \mid subject) \tag{4.7}$$

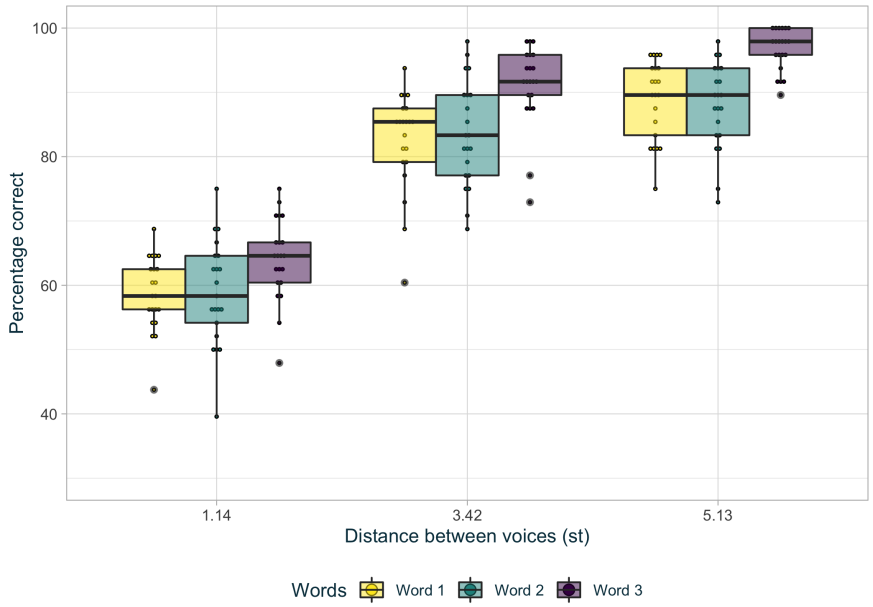

Figure 4.5 : Percentage of correct responses for each voice for every keyword.

Based on the likelihood ratio tests, the position of the keyword in the sentence had an effect on participants' scores $[\chi^2(2) = 112.85, p < .001]$ as well as the voice $[\chi^2(1) = 985.58, p < .001]$ and the interaction $[\chi^2(2) = 41.65, p < .001]$. Post-hoc analysis showed that participants obtained higher scores when the keywords were at the end of the story than at the beginning $[z = -7.08, p < .001]$ or in the middle of the story $[z = -6.96, p < .001]$, for the three voice conditions.

Error analysis

Figure 4.6 illustrates the error distribution and Equation 4.8 represents the final model of a top-down strategy modelling gLMM on binary (masker or extraneous) data from subset of answers where the participant did not respond the target keyword. Participants chose a keyword uttered by the masker voice over the extraneous keyword when distance between the target and the masker voices decreased $[\beta = -1.19, SE = 0.14; z = -8.29, p < .001]$.

$$errortype \sim voice + (voice \mid subject) \tag{4.8}$$

The ratio of masker errors to extraneous errors was higher than chance only when the distance between voices is 1.14 semitones (see Table 4.2). Chance was computed for each voice with a binomial test.

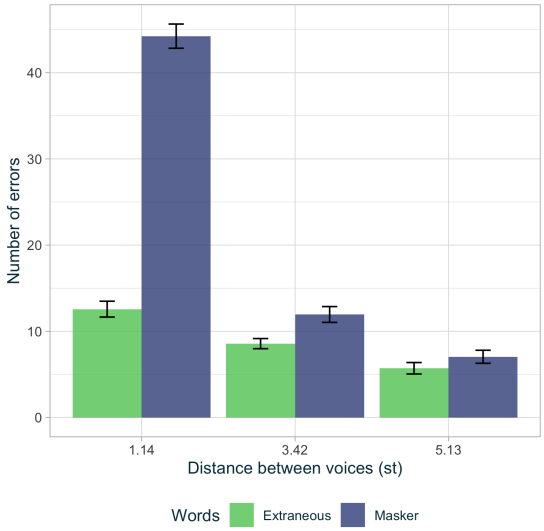

Figure 4.6 : Average number of errors for each condition. The bars represent the masker answers (in blue) and the extraneous answers (in green). The error bars are the standard error of the mean.

Table 4.2 Comparisons with chance level for the ratio of masker errors to extraneous errors.

Voice distance	Statistics	
	z	p
1.14	13.06	$< .001$
3.42	1.24	.22
5.13	0.09	.93

Semantic context

A gLMM was fitted on the binary data (target or masker) with the semantic context, the distance between voices and stimulus presentation. Equation 4.9 shows the final model with a top-down modelling:

$$score \sim voice + \Phi_k + (voice \mid subject) \tag{4.9}$$

As seen before, participants had better scores when the distance between voices increased $[\beta = 2.55, SE = 0.13; z = 1.87, p < .001]$ but in addition, their score also depended on the semantic context: higher semantic context yielded higher scores $[\beta = 1.44, SE = 0.25; z = 5.86, p < .001]$.

Comparison with behavioural experiments 1 and 2

To conclude the present section on behavioural results, participants's performance at the Long-SWoRD test are congruent to experiment 1 and experiment 2 presented in Chapter 3. Participants from this third study scored the same as participants in experiment 1 (see Figure 3.4; $[5.13 : t(41) = 0.48, p = 0.64]$) and participants in experiment 2 (see Figure 3.9; $[1.14 : t(49) = -1.91, p = 0.06; 3.42 : t(49) = -0.18, p = 0.86)$ under all three conditions (see Appendix A for more informations). The ratio of masker errors to extraneous errors was higher only for the most difficult voice condition, 1.14 st. In addition, participants also had better performances for the end-word and could also benefit from the semantic context. Finally, contrary to the experiment 2 from Chapter 3, when the distance between the two voices was 1.14 st, participants did use the semantic context to select the keywords.

4.4.2 Neural tracking results

Temporal Response Function

The average TRFs for target and masker reconstruction, per voice condition, are shown in figure Figure 4.7. Two late responses, usually identified as cognitive, can be observed. First, a prominent negativity was apparent for time lags between 280 and 540 ms ($N400_{TRF}$) for all the TRFs (see Table 4.3). The amplitude of this negativity was larger for the masker than the target when the distance between the two voices is 1.14 or 3.42 st. It is interesting to note the presence of this $N400_{TRF}$ marker in the most difficult condition, 1.14 st. As mentioned in the Section 4.4.1, the only difference between this present experiment and the two experiments from Chapter 3 is the use of semantic context by participants in the most adverse listening condition. The presence of the $N400_{TRF}$ marker strengthens the idea that participants could benefit from the semantic context to answer to the Long-SWoRD test. A second cognitive component can also be observed at a later latency (670-800 ms). A prominent negativity for the masker was apparent for voice conditions of 1.14 and 3.42 st while a prominent positivity for the target could be found when the distance between the voices is 1.14. st. These markers will be referred to as $N7_{TRF}$ and $P7_{TRF}$ for the remainder of this chapter.

In addition, a very early positivity (between 0 and 50/100 ms) was present for the target in the most difficult condition and for all the maskers. This led to

Table 4.3 : TRF Markers latency for each stream and condition (in ms).

Label	**Target**			**Masker**			**Difference**		
	1.14 st	*3.42 st*	*5.13 st*	*1.14 st*	*3.42 st*	*5.13 st*	*1.14 st*	*3.42 st*	*5.13 st*
"P1"	0-31	94-125		0-15	0-109	0-47		0-140	62-78
N1									
N4	281-391	344-547	296-532	281-578	234-578	328-513	391-500	356-484	
	500-547								
"P7"				687-750	672-813		687-703	687-781	
"N7"	719-750								

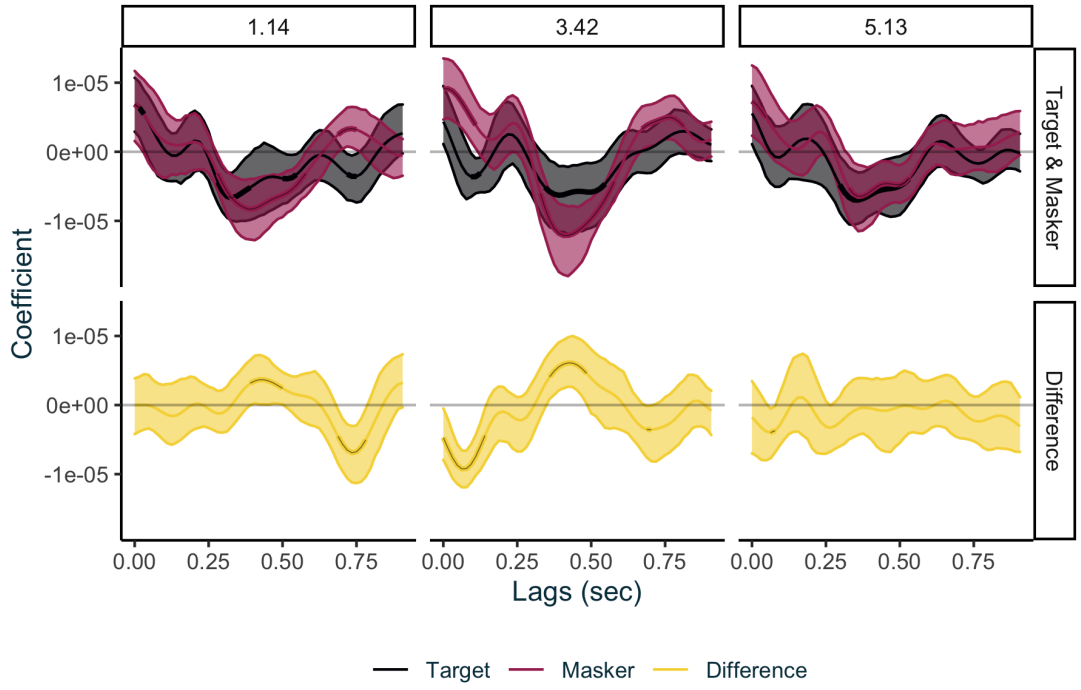

Figure 4.7 Temporal Response Function. Thick lines indicate a response significantly different from zero based on the confidence interval. Confidence intervals (95%) were obtained by bootstrapping the mean across subjects. Top row: TRF for the target stream (in black) and for the masker stream (in rose) for each condition. Bottom row: Difference between target and masker TRF.

a difference of amplitude at an early latency between the target TRF and the masker TRF when the distance between the two voices was 3.42 or 5.13 st. It is noteworthy that the difference in the latter condition is very tiny. The nature of this component is open to wide discussion since this marker seems to start before the latency is equal to zero. Hence, when the latency is null (t=0), TRFs coefficients are different from zero. As this is an early positive component, however, it will be labelled $P1_{TRF}$ for the rest of this chapter.

We analyse TRFs per region of interest (ROI) to better understand the presence of this non-zero component. It is noteworthy that it was not a source localization analysis but rather a basic localization analysis. Figure 4.8 represents the average TRFs for each condition . The frontal TRF was the averaged TRF computed with electrodes 'FP1', 'FP2', 'F3', 'F4' and 'Fz', the left parieto-temporal TRF with 'T7', 'TP9', 'P7', 'CP5', 'P3', 'PO9' and finally the right parieto-temporal TRF with 'T8', 'TP10', 'P8', 'CP6', 'P4', 'PO10'. It can be observed that the positive early response, $P1_{TRF}$, is found exclusively in the frontal area. It can also

be briefly noted that the differences between the target and the masker streams are the most salient in the left parieto-temporal area (see Appendix A for more details).

Stimuli -reconstruction evaluation (R)

Figure 4.9 shows the average correlation between the reconstructed envelope and the original envelope for each subject per condition. In this analysis, the reconstruction of the target envelope was performed with a decoder (TRF) trained on the original target envelope and the reconstruction of the masker envelope was performed with a decoder (TRF) trained on the original masker envelope (see Appendix A for more details). As the voices become more different from each other and the task becomes easier, the target reconstruction improves (higher R-value) $[R_{1.14} = 0.097; R_{3.42} = 0.114; R_{5.13} = 0.116]$ and the masker reconstruction worsens $[R_{1.14} = 0.074; R_{3.42} = 0.054; R_{5.13} = 0.05]$. All the reconstructions were above chance (multiple paired t-test between the original and the surrogate data with a false discovery rate correction).

A linear mixed model (LMM) was fitted on the (Fisher transformed) Pearson's correlation scores (R):

$$R \sim voice * decoder + (voice \mid subject) \tag{4.10}$$

There was a difference in the stimulus-reconstruction depending on the speech (masker vs. target) stream $[\chi^2(1) = 596, p < .001]$ but there was no effect of voice $[\chi^2(2) = 0.52, p > .05]$. On the other hand, the interaction was significant $[\chi^2(2) = 91.78, p < .001]$. Post hoc analysis showed that the target speech reconstruction was better than that of the masker for all conditions. In addition, the target envelope was better reconstructed for the two larger voice differences, 3.42 and 5.13 st, than for the smallest distance of 1.14 st $[t(20) = -4.56, p < .001; t(20) = -4.04, p < .001]$. The masker envelope was better reconstructed for a distance of 1.14 st than for both distances of 3.42 and 5.13 s $[t(20) = 4.93, p < .001; t(20) = 5.96, p < .001]$.

Analysing the R value resulting from each individual lag (see Figure 4.10), it appears that the stimulus-reconstruction became worse as the latency increased. However, there was a key moment around 200 ms: while the reconstruction of

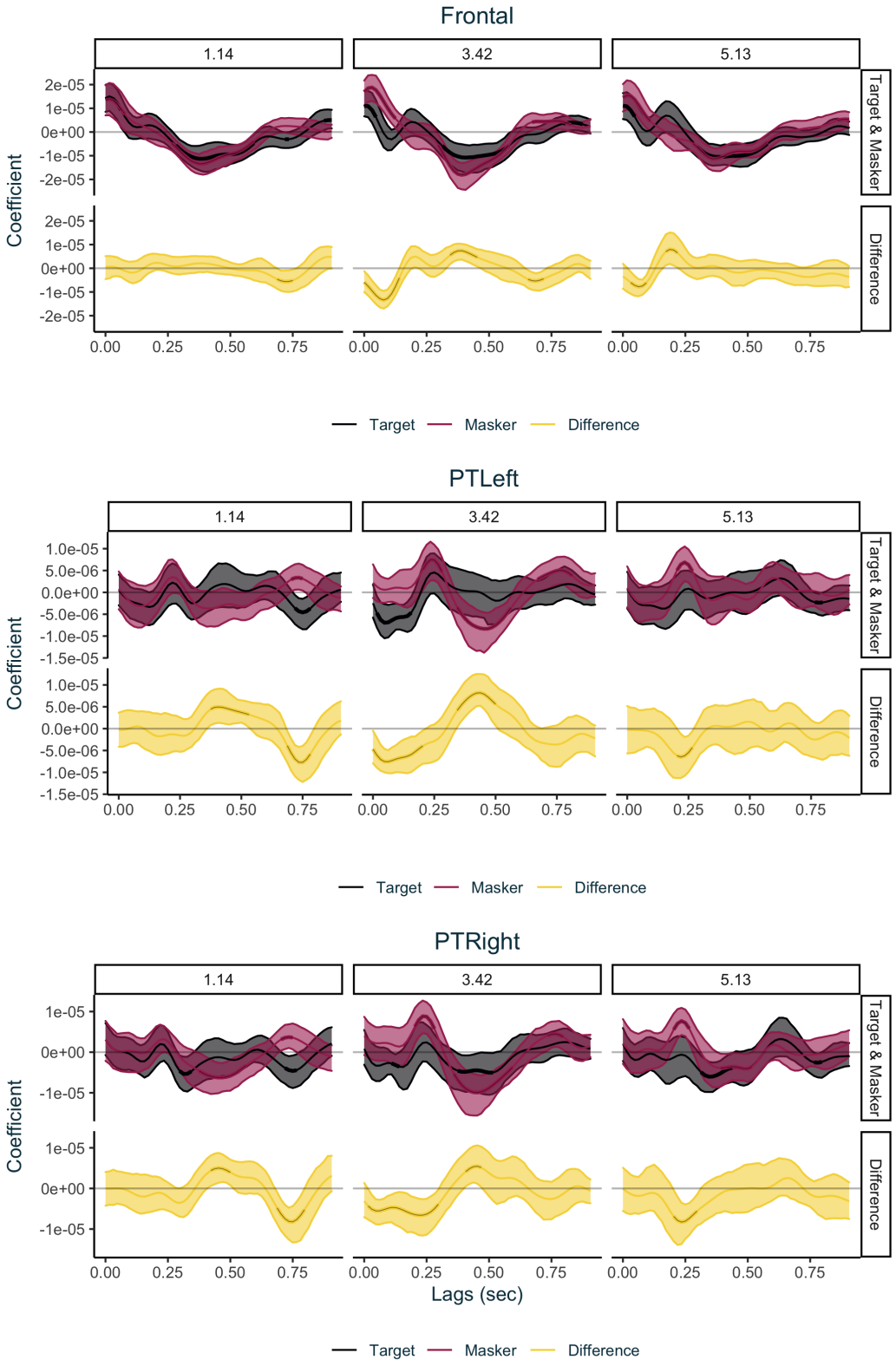

Figure 4.8 : Temporal Response Function for each ROI.

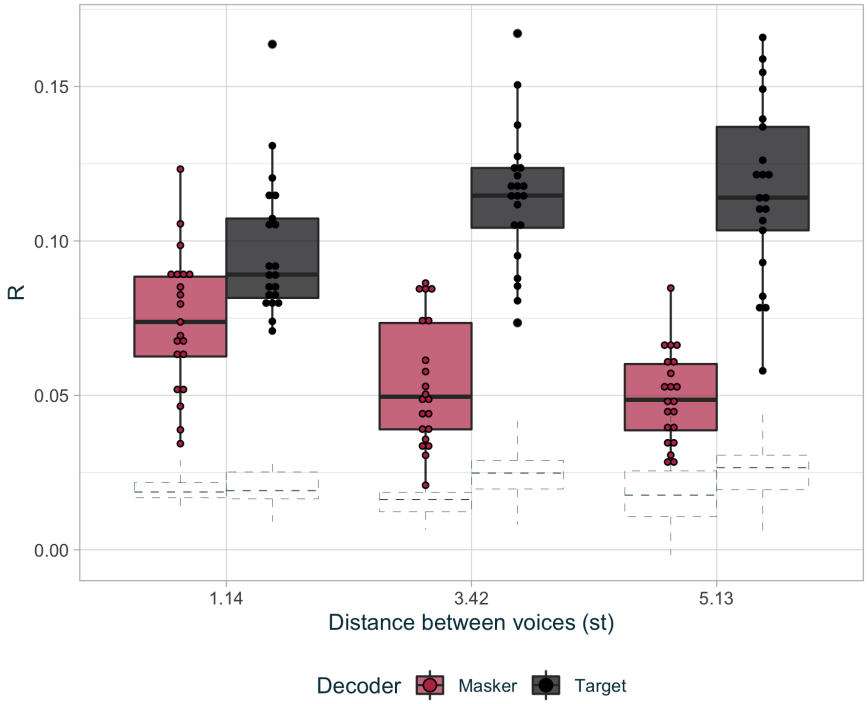

Figure 4.9 : Pearson's correlation between reconstructed envelope (using the TRF) and the original signal's envelope for each voice. The dots represent the scores for every participant in each condition for the target stream (in black) and for the masker stream (in rose). The hinges of the boxplot represent the first and the third quartile. The median is represented as a bar in each boxplot. The length of the whiskers is 1.5 interquartile range. The dashed boxplot indicates the surrogate data (level of chance).

the target is getting stronger, the reconstruction of the masker dropped abruptly. This effect is stronger when the distance between voices increases.

Auditory attention decoding

The auditory attention decoding (AAD) is the percentage of trials correctly classified by the decoders (Figure 4.11; see Appendix A for more details). When the TRF was trained with masker envelopes, the AAD was never above chance. When the TRF was trained with target envelopes, only 9, 17 and 18 participants out of 21 were above chance when the distance was, respectively, 1.14 st [$\mu_{1.14} = 60.12\%$] , 3.42 st [$\mu_{3.42} = 72.32\%$] and 5.13 st [$\mu_{5.13} = 76.59\%$] between voices.

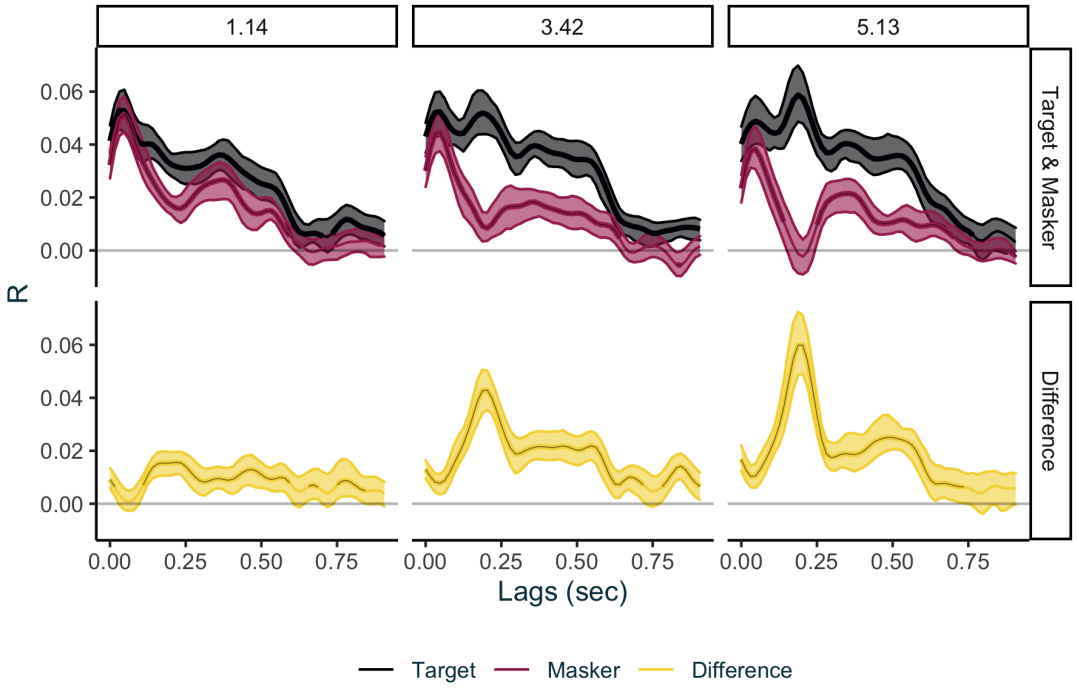

Figure 4.10 Evaluation of stimulus-reconstruction (R) per lag. Thick lines indicate a response significantly different from zero based on the confidence interval. Confidence intervals (95%) were obtained by bootstrapping the mean across subjects. Top row: R for the target stream (in black) and for the masker stream (in rose) for each condition. Bottom row: Difference between target and masker TRF.

A generalized linear mixed model (gLMM) was fitted on the binary (correct/incorrect) AAD.

$$AAD \sim voice * decoder + (voice \mid subject) \tag{4.11}$$

There was a difference in AAD depending on the speech (masker vs. target) stream [$\chi^2(1) = 470.66, p < .001$] but there was no voice effect. On the other hand, the interaction was significant [$\chi^2(2) = 79.61, p < .001$]. Post hoc analysis showed that the AAD was better when the decoder is trained by target envelopes than by masker envelopes for all the conditions [$AAD_{1.14} : z = -5.36, p < .001; AAD_{3.42} : z = -14.2, p < .001; AAD_{5.13} : z = -17, p < .001$]. In addition, the target AAD was better for the distances 3.42 and 5.13 st than for a distance of 1.14 st [$z = -5.79, p < .001; z = -7.95, p < .001$]. The target AAD was also better with a distance of 5.13 st between voices than with a distance 3.42 st between voices [$z = -2.19, p < .05$]. The masker AAD was worse for both distances of 3.42 and 5.13 st than for a distance of 1.14 st [$z = 3.27, p < .01; z = 4.09, p < .001$].

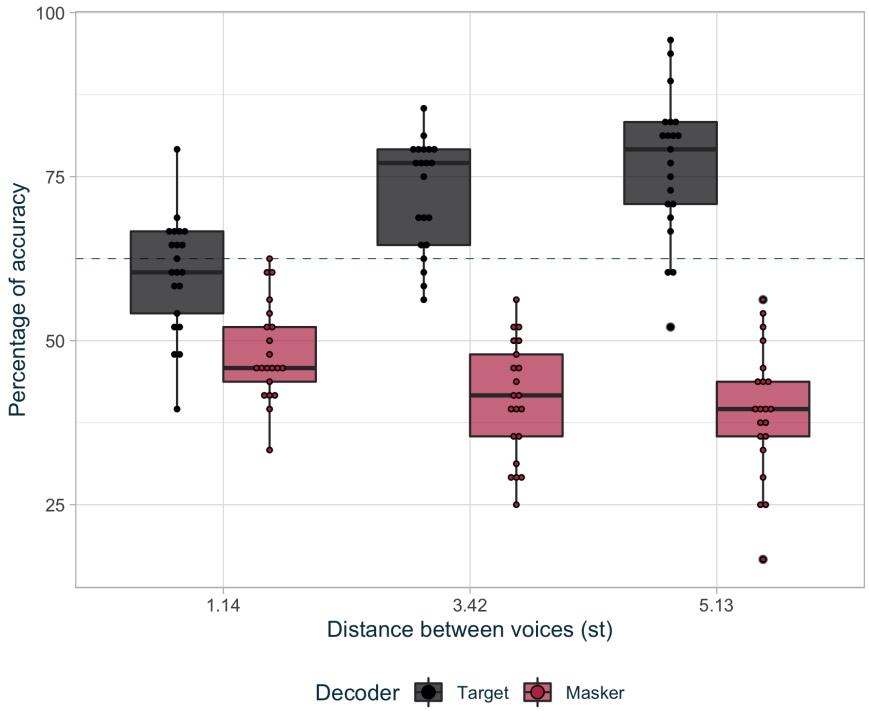

Figure 4.11 : Auditory attention decoding (AAD). The dots represent the percentage for every participant in each condition for the target decoder (in black) and for the masker decoder (in rose). The hinges of the boxplot represent the first and the third quartile. The median is represented as a bar in each boxplot. The length of the whiskers is 1.5 interquartile range. The dashed line (62.5%) indicates the level at which performance is significantly greater than chance based on a binomial test at the 5% significance level.

Behavioural vs. Neural

In order to investigate the estimation of the intelligibility though cortical tracking measures, we compared behavioural and neural results. Figure 4.12 shows the common relationship between behavioural scores (Long-SWoRD) and neural measures [Stimulus-reconstruction Evaluation (R) and Auditory Attention Decoding (AAD)]. However, these two measures can be explained by the difficulty of the task (the distance between the target and masker voices). Thus, the confounding variable condition needed to be partialled out in order to analyse the relationship between behavioural performance and neural measures *within* voice conditions. To do so, the neural and behavioural measures were centre-reduced per condition (see Figure 4.13). Overall, the relationship between the two neural measurements and the Long-SWoRD test was weak for both the target stream and the masker stream.

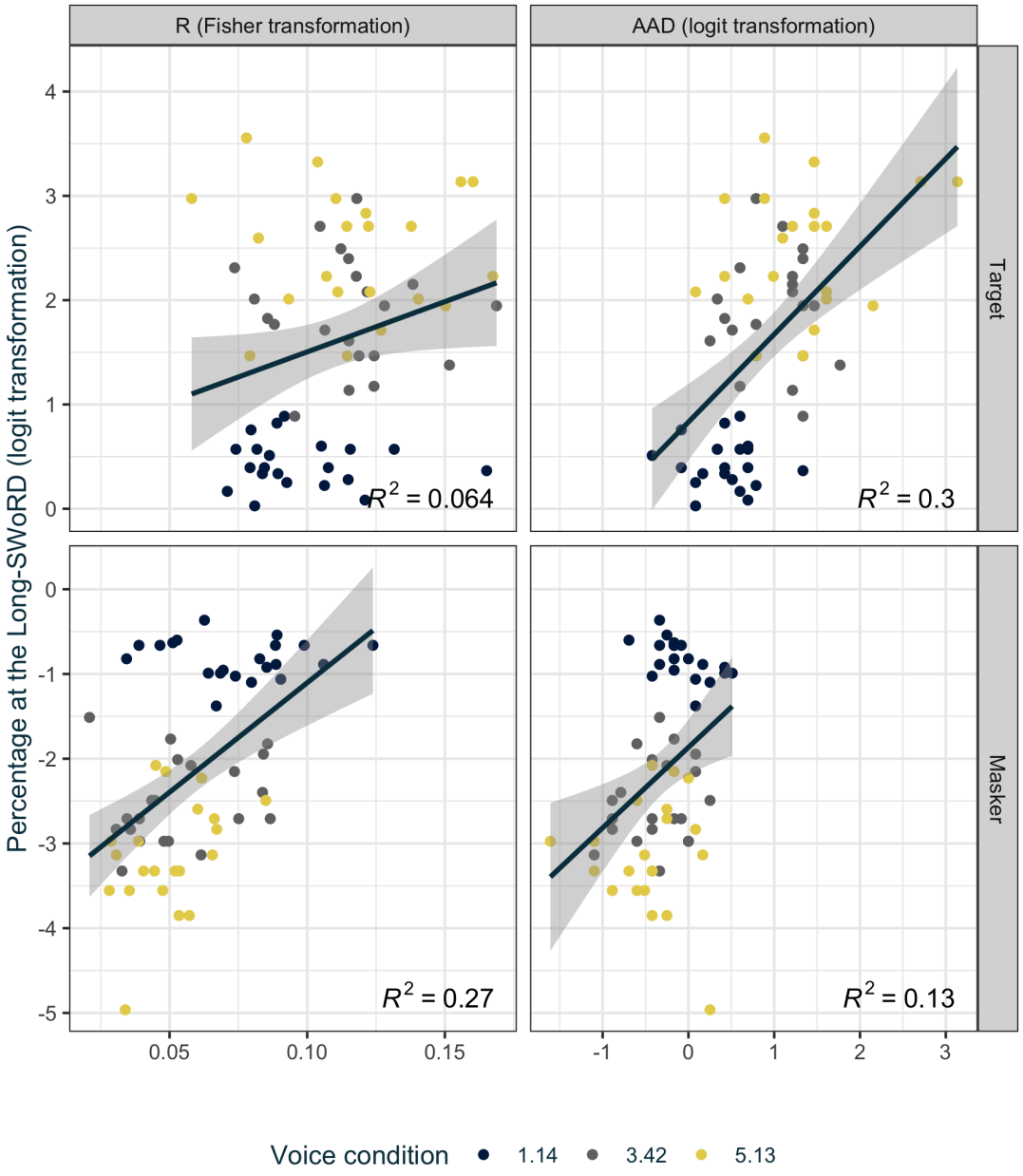

Figure 4.12 : Long-SWoRD scores as a function of neural measures [Stimulus-reconstruction evaluation (R) on the left and auditory attention decoding (AAD) on the right] for the Target and the Masker measures. Each subject is represented by three dots (one for each condition). Predictions from the linear model are shown with 95% confidence intervals

.

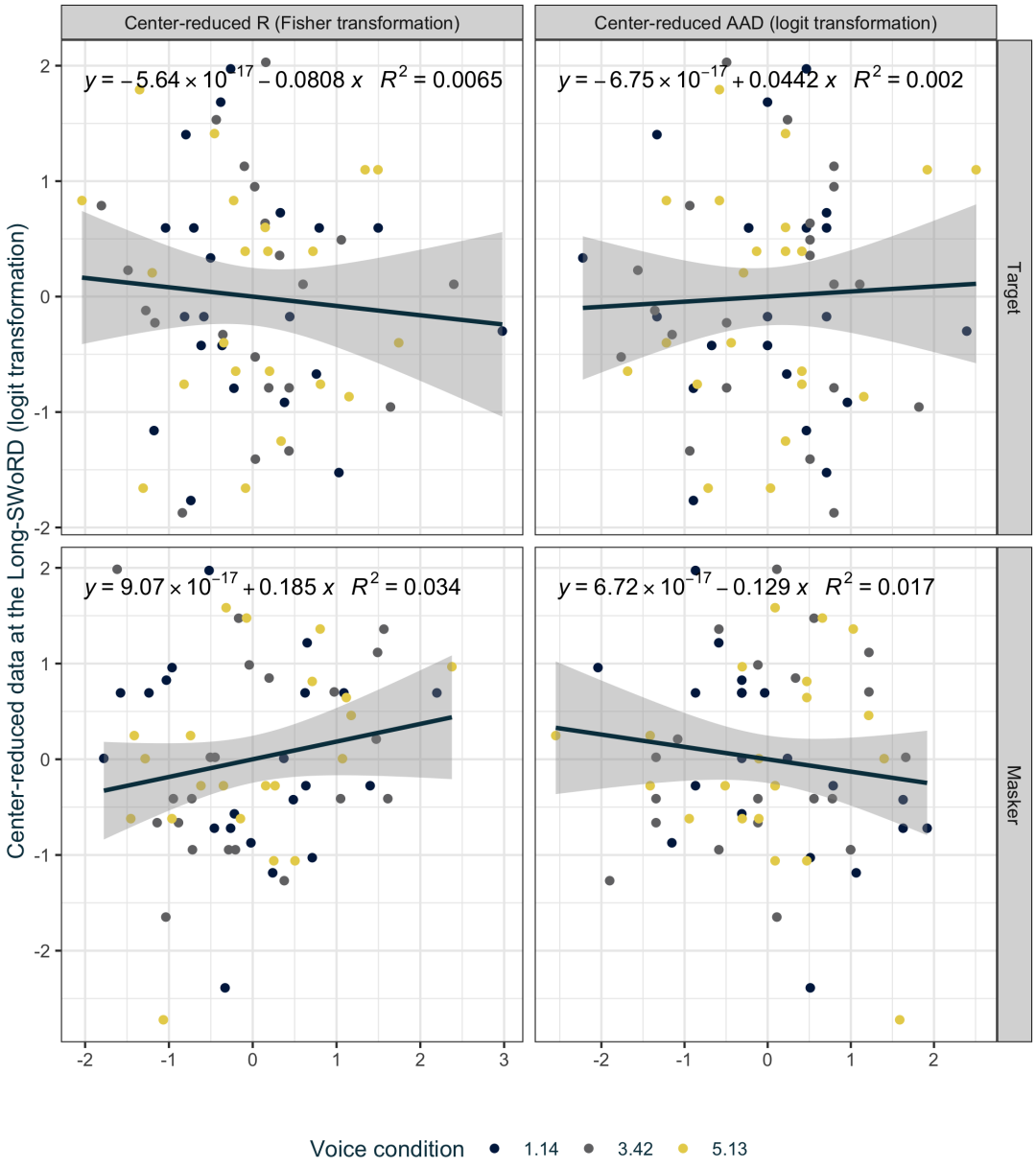

Figure 4.13 : Long-SWoRD scores as a function of neural measures [Stimulus-reconstruction evaluation (R) on the left and auditory attention decoding (AAD) on the right] for the Target and the Masker measures. The data were centre-reduced per condition on the Fisher transformed R data and the logit transformed of the Long-SWoRD test and the AAD data. Each subject is represented by three dots (one for each condition). Predictions from the linear model are shown with 95% confidence intervals

.

4.4.3 Build-up effect

Target and masker

This section investigates the segregation of the mixture in two streams *over time*, also called the *build-up effect*. For this analysis, a sliding window was used to reconstruct speech streams instead of the entire trial. Capturing information at three key moments (beginning, middle and end of the story) independently was important to analyse the evolution of the build-up effect. Thus, the duration of the sliding window is one third of the duration of the Long-SWoRD test shortest story (11 s). To be more precise, the first 3.66 s of each trial were used to train the TRFs and then to reconstruct the first 3.66 s of each trial. As in previous analyses, this process was applied to the target stream and to the masker stream. The analysis was then repeated with the same window size by steps of 0.1 s. Finally, the analyses were performed only on the first 11 s of each story (the duration of the shortest story).

Since the stimulus-reconstruction might not be linear, a generalized additive model (GAM) was fitted for each voice and speech streams on the stimulus-reconstruction evaluation (R; see Equation 4.12).

$$R \sim s(time, by = structure) + structure + s(time, subject, bs = fs, m = 1, by = structure) \quad (4.12)$$

where *structure* represents each reconstruction evaluation (6 levels: 3 per voice multiplied 2 for target and masker) and *time* is the last point of the sliding windows. GAM statistics (see Table 4.4) confirm that regressions were non-linear (except for the target speech is the most difficult condition.). Moreover, the reconstruction evaluation was changing over time for the target in the easiest condition and for the masker in the most difficult condition.

Interpreting GAM is not easy because the statistics do not tell exactly where the curves are different from zero, nor the amplitude of that difference. Hence, visualization of the model is required. Figure 4.14 (top row) shows that the reconstruction of the target improved over time in easy cocktail party situations. In contrast, the reconstruction of the target in difficult listening conditions remained stationary. Reconstruction of the masker tended to fall off slowly over time under easy voice conditions. When the distance between voices was 1.14 st, the masker reconstruction oscillated over time.

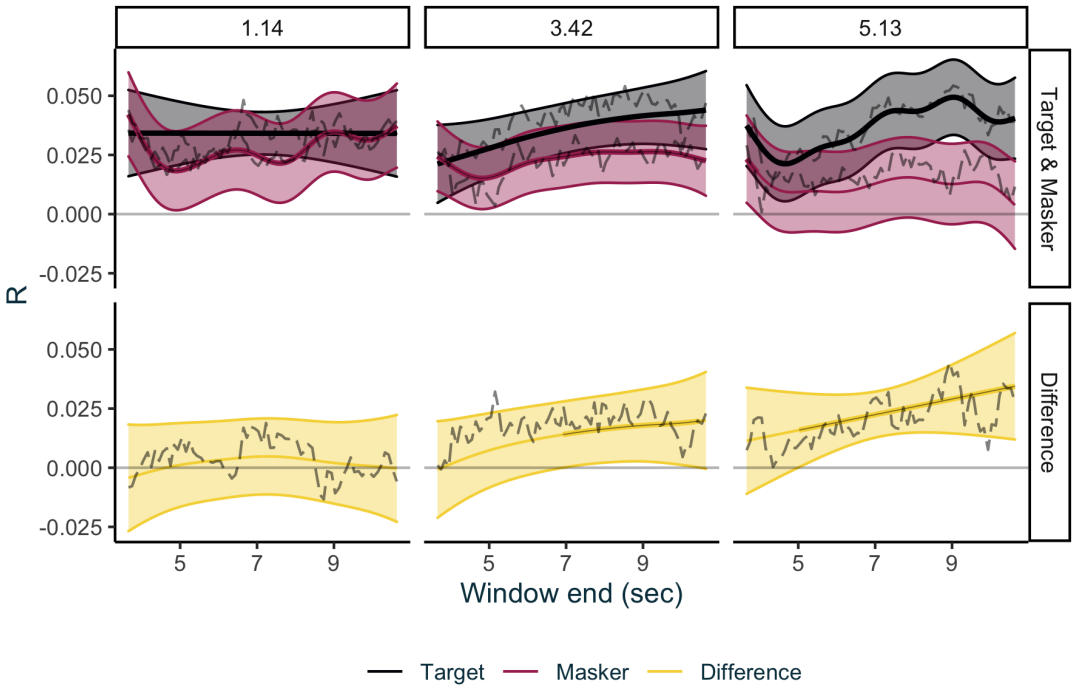

Figure 4.14 : Stimulus-reconstruction (R) over time per condition for the target stream (in black) and the masker stream (in rose). The difference between the target and the masker streams is shown at the bottom row in yellow. Thick lines indicate a response different from zero with 95% confidence intervals. The dashed lines indicate the raw data averaged (before the smoothing of the GAM).

Table 4.4 GAM smooth term parameters for each stream reconstruction per condition. Each line shows if the non-linear regression associated to the smooth term is significantly different over time. The effective degrees of freedom for the model parameter (edfs) provide an estimation of the wiggliness of the smooth. An edf of 1 means that the smooth is linear. The higher the edf is, the more complex the smooth is.

Smooth terms:	*Edf*	*F*	*p*
$s(time) : Target - 1.14$	1.0	0.0	0.99
$s(time) : Target - 3.42$	1.78	1.88	0.23
$s(time) : Target - 5.13$	7.06	2.09	$< .05$
$s(time) : Mask - 1.14$	7.1	4.0	$< .001$
$s(time) : Mask - 3.42$	4.79	1.01	0.31
$s(time) : Mask - 5.13$	5.72	0.98	0.43
$s(time) : Mixture - 1.14$	7.64	3.28	$< .001$
$s(time) : Mixture - 3.42$	1.0	2.44	0.99
$s(time) : Mixture - 5.13$	7.69	3.81	$< .001$

The difference between the target and the masker reconstruction was also modelled according to Equation 4.12 and is displayed in the bottom row of Figure 4.14. There was no difference over time between the target and the masker for the most difficult voice condition (1.14 st). However, the target reconstruction was better than the masker reconstruction from 6.92 s onwards for the intermediate difficulty and from 5.02 s for the easiest condition.

Mixture and target

As already mentioned, the build-up is the time required to separate a mixture into two streams. To get a complete picture, the reconstruction of the mixture was also analysed following the same build-up analysis. Then, the reconstruction of the mixture with a GAM for each voice with the same Equation 4.12. Surprisingly, the reconstruction of the mixture with a distance of 3.42 st followed a straight line, unlike the reconstruction of the mixture under the other two conditions which change significantly over time (see Table 4.4).

In addition, the three mixture reconstructions are shown in Figure 4.15 along with the difference between the target and the mixture neural tracking accuracy. Mixture reconstructions behaved identically to target reconstructions in condition where there was a distance of 5.13 st: the reconstruction evolved over time with a slight drop at first around 4 s and then an improvement. The two reconstructions

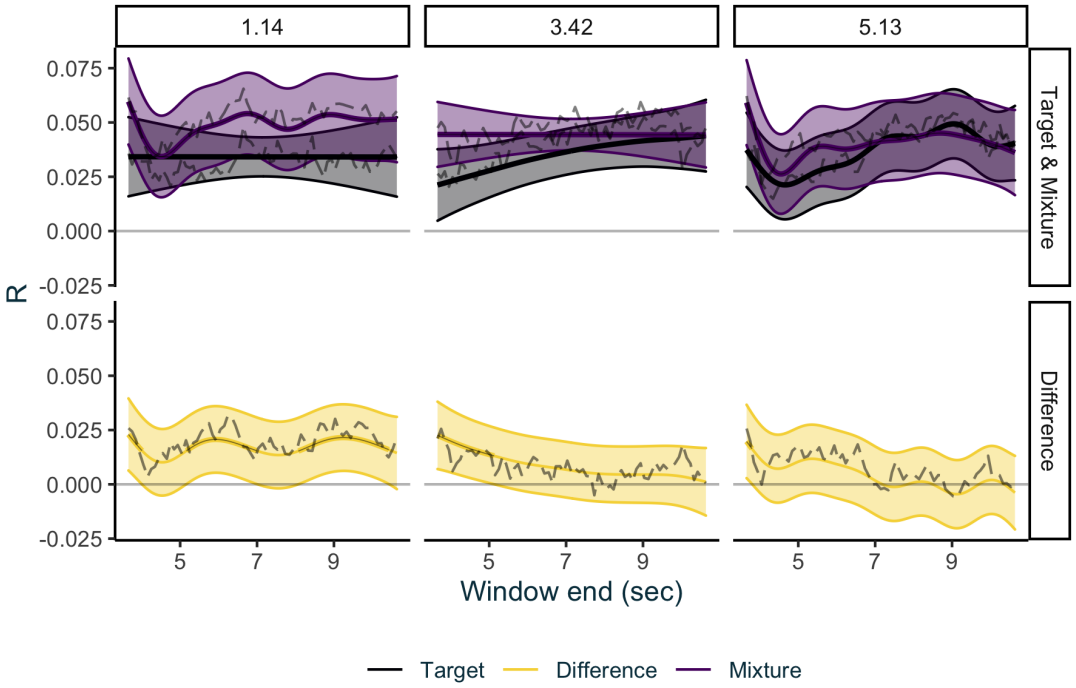

Figure 4.15 : Stimulus-reconstruction (R) over time per condition for the target stream (in black) and the mixture stream (in purple). The difference between the target and the mixture streams is shown at the bottom row in yellow. Thick lines indicate a response different from zero with 95% confidence intervals. The dashed lines indicate the raw data averaged (before the smoothing of the GAM).

were however briefly different before 3.8 s. When there was a distance of 3.42 st between the voices, the two reconstructions were linear but with different slopes and intercepts, resulting in a better reconstruction for the mixture before 5.13 s. When voices were separated by 1.14 st, the linear target reconstruction reached intermittently the mixture reconstruction. It is noteworthy that the masker reconstruction never reached the mixture reconstruction (see Appendix A for more details).

Finally, a brief drop appears in the representation of the mixture around 4 s. In the voice condition 3.42 st, this drop is not visible in the GAM modelling but appears in the raw data. One possible interpretation of this drop is that this would be the end of a control period during which attention is highly sustained by participants. At the end of this period, either participants manage to reach an optimal representation of the target on the basis of acoustic cues (such as in easy condition), or a change of strategy has to be made (such as in the most difficult condition). Under the intermediate condition, the participants could maintain this

active sensory decomposition but require more concentration. Investigations are needed to determine what happens before these 3.66 s but these future analyses would be limited by the constraint of the size of the data required for TRFs (see Chapter 5).

4.5 Discussion and conclusion

4.5.1 Comparison with the literature

In general, neural tracking results are consistent with the literature and our hypotheses. The target is better represented than the masker and the difference between both voices increases while the difficulty of the task is reduced. This difference seems particularly enhanced around the critical period of 200 ms, in line with previous studies (*e.g.* O'Sullivan et al., 2015). Auditory attention decoding (AAD) is above chance under easy conditions although results are below those in the literature (see Chapter 5). We can also observe the systematic presence of an $N400_{TRF}$ suggesting a semantic treatment of the stories. Broderick et al. (2018) observed a greater amplitude of the $N400_{TRF}$ for the target stream than for the masker stream. In our experiment, however, the $N400_{TRF}$ is more prominent for the masker story than the target story under adverse listening conditions. Surprisingly, a very early positive component can also be observed, mainly modulated by the frontal region for all the masker streams as well as for the target stream in the most difficult condition. In a cross-correlation study, Kong, Mullangi, and Ding (2014) observed a prominent P1 response for the masker speech streams. The authors hypothesized that the interaction between a P1 response and attention-modulated N1 and P2 responses might enhance the P1 component. Hence a weaker N1 amplitude might reinforce a P1 response. Conversely, the presence of an N1, as is the case for the stream target in the intermediate condition, indicates modulation of attention.

Previous studies (Decruy et al., 2019; Lesenfants et al., 2019; Vanthornhout et al., 2018) reported a relationship between neural tracking and intelligibility. We also observed a correlation between neural tracking and behavioural performance in the Long-SWoRD test but these measures were modulated mainly by the difficulty of the task. Hence, at equal difficulty, there is no link between these two measures. Thus, good neural tracking for a participant does not necessarily lead to good behavioural performance on the Long-SWoRD. However, top-down inputs

such as cognitive load can influence neural tracking as Hjortkjaer et al. (2018) pointed out. It is therefore very likely that cognitive resources modulates the link between intelligibility and speech tracking.

4.5.2 Resolution strategies

Placed in perspective with the behavioural outcomes of the Long-SWoRD, neural tracking results and build-up analysis seem to indicate three different resolution patterns in the Long-SWoRD test.

In the easy voice condition, 5.13 st, participants' behavioural scores and neural tracking measures are optimal. In this context, the differences in treatment between the two voices seem to occur very quickly, suggesting a main sensory treatment of the stimuli. In addition, as suggested by the presence of a $N400_{TRF}$ in TRFs, the participants seem to analyse the semantic information of the target and the semantic information of the masker in an analogous way, which could be a method to ensure a good score in the Long-SWoRD test: by knowing the semantic field of the masker, participants decrease the risk of errors.

In the intermediate condition, the participants make slightly more errors than in the easy condition, but these errors are not attributed to the selection of the word masker. As suggested in Chapter 3, in this configuration, it is likely that participant errors are due to distraction or increased cognitive load. Neural differences can also be observed between the easy and the intermediate voice conditions with AAD: there is a lower AAD in the intermediate condition than in the easy condition. However, we observe, in the intermediate condition, an identical stimulus-reconstruction to the easy condition even though the intermediate condition is more demanding. This observation is in contradiction to Hjortkjaer et al.'s results (2018) who showed that when the task is more demanding, cognitive resources are reallocated leading to a weaker stimulus-reconstruction. A possible explanation for this contrast is the nature of the task difference between Hjortkjaer et al.'s study (2018), an n-back task with noise and our present study. Finally, TRFs suggest that the semantic information of the target and the masker streams are analysed by participants. TRFs also highlight differences between the target and masker streams at early latencies (sensory treatment) and later latencies (semantic treatment). Thus, in the intermediate condition, participants would rely on a strategy mainly dominated by sensory information, and are complemented

with linguistic cues to ensure an identical reconstruction of the streams as in the easy condition.

In the most difficult condition, the participants still manage to complete the task, but they make mistakes that can be attributed to listening, at least partially, to the masker. Unsurprisingly, the stimulus-reconstruction and AAD are not as good as in the other two conditions. TRFs highlight differences in semantic and very late components ($P7_{TRF}$, $N7_{TRF}$). In addition, the build-up analysis indicated that it does not appear to a have a clear segregation of the voices. One explanation might be that the target and the masker build up over time but may be partially or completely reset by a sudden change in the properties of the speech signal or by switches in attention (B. C. J. Moore & Gockel, 2012). In conclusion, the success of Long-SWoRD in this difficult condition seems overall to be explained with a purely cognitive strategy.

4.5.3 Build-up effect

Under conditions where early neural responses are observed and where therefore sensory mechanisms are implemented, there is a difference in neural tracking over time for the target and the masker. According to Bregman (1994), the duration of the build-up relies on the salience between the two tones and is supported by our results that describe a faster separation for the easiest condition. This build-up seems to be the product of two phenomena. On the one hand, the masker is gradually inhibited. And on the other hand, the representation of the target is gradually enhanced. It appears that once the target is identified, its representation would be similar with the representation of the mixture, which is a representation of the holistic scene.

4.5.4 Conclusion

In conclusion, as well as performance in the Long-SWoRD test, neural tracking measures are modulated by the difficulty of the task. In addition, TRFs markers can be identified and informed on strategies used by the listeners. Until now, behavioural and neural measures have been used in parallel. The next chapter will investigate the combination of behavioural and neural measures, and more specifically, how behavioural information can enhance the construction of the decoder speakers.

Chapter 5

Behavioural enhancement of the neural tracking

5.1 Introduction

Building an attention decoder from EEG recordings is a time consuming endeavour. In their initial study, O'Sullivan et al. (2015) used 30 minutes of recordings per participant. This seriously hinders the usability of neural tracking both for research purposes as well as for clinical applications. The possibility of improving hearing-impaired technology is one of the major motivations of this enthusiasm (see Chapter 1 for more information). Numerous papers have been published over the last five years with the aim of improving the performance of *"neural tracking"* , and sometimes more specifically the *"Auditory Attention Decoding" (AAD)* method (see Chapter 4 for more information). In general, research has focused on two aspects of neural tracking enhancement: (1) the robustness of the Temporal Response Function (TRF), and (2) the parameters of the stimuli used. These two aspects are briefly presented in the following sections. Other approaches, such as deep neural networks, that are related to the TRF technique have also benefited from improvements over the years, for similar reasons (see Chapter 1 for more information). For the sake of brevity, these approaches are not detailed here and this chapter is limited to research on the TRF literature.

5.1.1 Robustness of the neural tracking

In general, two main objectives are pursued with the study of neural tracking. First of all, a clinical application that aims to improve hearing-impaired technologies. The ultimate goal is to determine the voice of the speaker on which the listener is focusing on in order to, for example, increase the voice of the target speaker while attenuating the background sounds. The challenge is therefore to obtain the best AAD, but with as little data as possible in order to be able to move towards real-time decoding. In this context, the robustness of neural tracking measures tests the limits of a certain level of proficiency for the TRF with the minimal amount of recordings. The second main objective pursued with the study of neural tracking is a better understanding of the perceptual and cognitive mechanisms involved in cocktail party situations. Here, the best neural representation of the speakers (stimulus-reconstruction) is aimed for. In addition, TRFs must also be stable and identical over trials since they can provide feedback on the strategies implemented by the listeners (see Chapter 4). In this context, robustness can be defined as the most accurate measures of stimulus-reconstruction and TRFs.

Several methodological aspects, both practical and mathematical, were tested for robustness by the researchers. Among the most practical aspects was Mirkovic et al.'s study 2015 that examined the number of electrodes required to obtain robust AAD. The results suggest that 25 electrodes are needed for optimal performance. Beyond this number, no improvement in the performance of the decoding process was observed. The authors also examined the amount of data required to perform robust TRFs. Attention decoding performance became better than chance when the TRFs where trained from 15 one-minute trials or more. With relation to trial duration, O'Sullivan et al. (2015) observed that 30 trials of 10 s produced an AAD above chance among 75% of the subjects. Fuglsang, Dau, and Hjortkjær (2017) also found that 25 electrodes and 10 s trials were enough to perform AAD above chance, even in noisier listening environments. Finally, for more computational aspects, when TRFs are trained and tested separately for each subject ("Subject-Specific" approach), it produces better results than when TRFs are trained on all subjects (*e.g.* O'Sullivan et al., 2015)

TRF is a mapping model that correlates an acoustic stimulus with its cortical response through linear regression. The very first papers only used the "Ordinary Least Squares" (OLS) function to solve the regression (O'Sullivan et al., 2015), then a ridge regularization was quickly associated with the Matlab toolbox proposed

by Crosse et al. (2016) to allow performance optimization. More recently, Wong et al. (2018) compared a series of regularization methods for both *"Backward"* and *"Forward"* models. Both models refer to the direction of regression. The backward model, as represented by Figure 5.1, maps from neural data to acoustic features while the forward model, conversely, maps from the acoustic features to the neural data. Wong et al. (2018) found that regularization methods such as the ridge regression, the low-rand approximation regression or the shrinkage regression, provide higher neural tracking measures than OLS regression for backward models. In addition, the authors found that backward models systematically outperformed the forward models in terms of regression and AAD. In general, the better performance of backward modelling can be explained by the additional information it contains compared to forward modelling. The backward model allows a multivariate analysis of all the electrodes at the same time while the forward model analyses channel by channel and is therefore blind to the data contained in the other channels (Crosse et al., 2016). However, one shortcoming of the backward model is that TRF coefficients cannot directly be interpreted as a measure of stimulus-related neural activity unless using the inversion procedure described by Haufe et al. (2014). In summary and in a schematic way, studies that have focused on the interpretation of TRF coefficients have tended to use forward models, whereas studies that investigated stimulus-reconstruction and AAD have opted instead for a backward approach with a regularization parameter.

Finally, the AAD in cocktail party situations with a forward approach is complicated because each pair of correlations per reconstructed electrode signal need to be evaluated. Usually, a support vector machine (SVM)[1] is required to perform such an evaluation. In contrast, a backward approach simply compares two stimulus-reconstructions and determines the largest correlation and, hence the best reconstruction. This difference can be explained by the traditional usage of acoustic representations that contain only one dimension as is the case with envelopes. In the following section, the different types of acoustic representations used for TRFs will be briefly presented.

[1] *Support Vector Machine* or *SVM* is a machine learning algorithm that analyzes data for classification and regression analysis

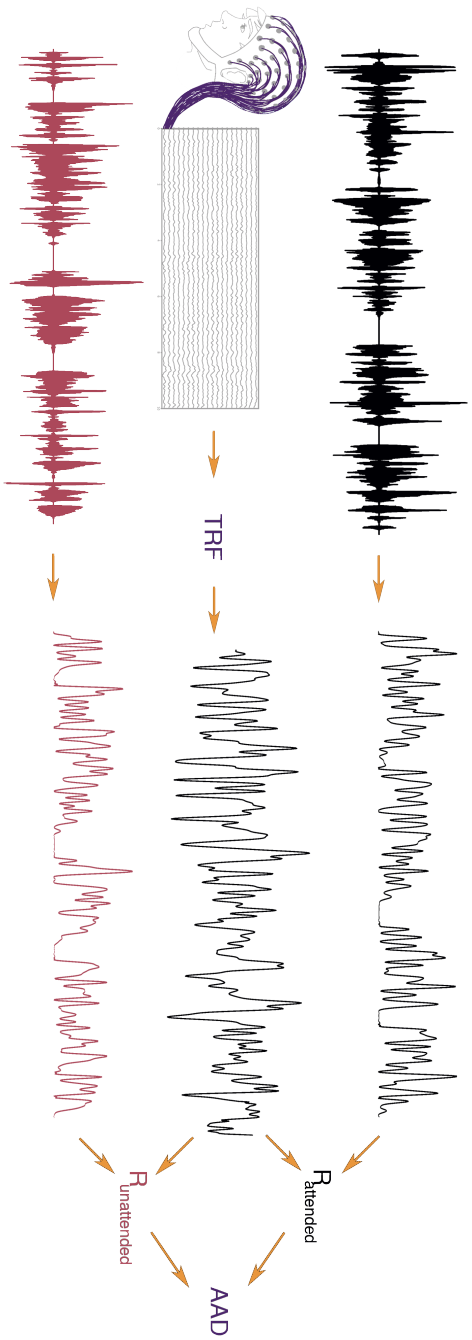

Figure 5.1 : Backward TRF approach. Neural data are regressed to predict the speech reconstruction. Correlations are then used to assess this prediction with the attended speech (in black; $R_{attended}$) and with the unattended speech (in red; $R_{unattended}$). AAD is achieved through the comparison of $R_{attended}$ and $R_{unattended}$.

5.1.2 Stimuli representations

Two representations have mainly been used to describe the stimuli in neural tracking studies, namely the spectrogram and the envelope (also known as temporal envelope or amplitude envelope). Di Liberto et al. (2015) trained forward TRFs with both representations associated with additional phonetic information. They showed that the relationship between continuous speech and neural activity is best described when the speech is represented with spectro-temporal and phonetic information. In addition, the speech reconstruction was better with the spectrogram representation than with the envelope. Apart from methodological considerations, this study therefore supported the idea that phonetic features are neurally encoded and can be tracked in continuous speech. Along the same lines, Drennan and Lalor (2019) have shown that the inclusion of information about signal intensity also leads to better reconstruction.

As mentioned in the previous section, the use of the envelope is widely preferred in backward modelling in order to facilitate AAD. However, there are several ways to extract an envelope from a signal. Biesmans, Das, Francart, and Bertrand (2017) were interested in the possible impact that the envelope extraction method could have on AAD performance. The results of this comparative study show that envelopes extracted on subband signals performed better than broadband-extracted envelopes. In addition, they have the advantage of mimicking the physiology of the auditory periphery. Finally, the inclusion of non-linear power law amplitude compression yielded the best performance and hence, the best neural representation of the speech stream.

5.2 Rationale

Enhancement of neural tracking, analysed with TRFs, has been approached in previous studies both with computational considerations and with different representations of the stimuli. It has already been pointed out in Chapter 1 that most studies on TRFs use long stimuli - at least a few tens of seconds and, in many of these papers, several minutes. Studies (*e.g.* Fuglsang et al., 2017; Mirkovic et al., 2015) have shown that it is possible to reduce the duration of stimuli to some extent, however, in practice researchers continue to employ long stimuli. The premise that listeners are able to maintain their attention focused on the target speech stream for relatively long periods of time is a shortcoming of studies

with long stimuli. Our own experience as participants in such tasks, and informal reports from others, strongly suggest that this assumption may not be warranted. In everyday-life situations, listeners are unlikely to maintain undivided attention on a single talker, and instead, can switch rapidly between different voices.

To mitigate this issue, some investigators have made attempts to control for such attentional-shift effects by supplementing it with a behavioural task. For instance, O'Sullivan et al. (2015) asked their listeners multiple-choice questions following every 1-minute stimulus, to check that the listener had been paying attention to the target story. One limitation of this approach, however, is that listeners may have been able to answer such questions correctly, even if they did not always pay close attention to the target story. In addition, this type of questionnaire raises the issue of intelligibility and comprehension measures raised in Chapter 2. Crosse et al. (2015) asked participants to press a button whenever they were listening to the target voice. However, firstly, it is possible that listeners were unable to track precisely the wanderings of their attention with their button presses, as they were listening to the story; secondly, asking listeners to press buttons according to their attention while they are listening introduces a secondary task, which may perturb performance of the primary, selective-attention task. It can be noted that studies have also worked on more computational aspects with Bayesian or state space modelling approaches in order to achieve AAD in real time (*e.g.* Akram et al., 2016, 2017; Miran et al., 2018).

The question of attentional switching between two auditory streams can hardly be discussed without first looking at the separation of these two auditory streams (see Chapter 1 for more information). And as mentioned in Chapter 3, spatial cues and voice cues are important for auditory speech segregation. Recently, two studies have investigated attentional switching with spatial cues. Bednar and Lalor (2020) showed that it was possible to reconstruct, with TRFs, the trajectory of attended and unattended moving sound sources. In the second paper, carried by Teoh and Lalor (2019), participants had to focus on a target voice while both talkers (target and masker) were instantaneously alternated between the left and right ears. The authors showed that it was possible to significantly improve AAD accuracy with the inclusion of spatial information even in 4 s stimuli.

The *Long-SWoRD* test was designed to provide experimenters with a means of inferring fluctuations in auditory selective attention. In this chapter, participants' answers from the experiment presented in Chapter 4 are used to infer, retrospectively, when they were listening to the target, or to the masker. The difficulty of

the task and hence the likelihood of an increase in attentional switch occur during the course of the stories was modulated with the similarity between target and masker voices. By combining the subjects' responses with different parameters, behavioural stimuli were created in order to best represent the actual attended speech and, therefore, should have a better reconstruction evaluation.

5.3 Methods

5.3.1 Participants, stimuli and procedure

Twenty-one participants undertook the *Long-SWoRD test* coupled with the audio stimuli from the audiobook *Le Charme discret de l'intestin* [The Inside Story of Our Body's Most Underrated Organ] (Enders et al., 2016) (see Chapter 2) while neural data were recorded with an EEG. The target and the masker story were presented diotically with a distance of 1.14 semitones (st), 3.42 st and 5.13 st. The same neural and behavioural data set, presented in Chapter 4, was used for the analyses in this Chapter.

5.3.2 Data acquisition and signal processing

Electroencephalography data were recorded using ActiCap (Brain Products, Munich, Germany) with a setup of 31 channels at a sampling rate of 1000 Hz. EEG data were then band-pass filtered between 2 Hz and 8 Hz. Finally, to decrease processing time, EEG data were downsampled to 64 Hz.

The stimulus speech envelope was extracted with a gammatone filterbank (Søndergaard & Majdak, 2013; Søndergaard et al., 2012) followed by a power law according to Biesmans et al. (2017). For more precision, the gammatone filter was used with 28 subband centred on frequencies from 50 Hz until 5000 Hz, equally spaced on the ERB scale. The 28 subbands envelopes were extracted by taking the absolute value of sample and then raising it to the power of 0.6. An unique envelope was then computed by averaging all the 28 envelopes. The speech envelope was then downsampled to 64 Hz and low-pass filtered below 8 Hz.

5.3.3 Behavioural stimuli

Behavioural stimuli estimate the actual attended stream, switching between target and masker streams, that participants were actually listening to. Participants' responses provide information at three key moments in the story (beginning, middle and ending keywords). It is important to note that there are not just three keywords that provide information, but six: three target and three masker keywords. For each keyword, the subjects cannot choose both target and masker, but they are faced with a choice. Thus, if a participant selects the masker keyword in their response, it is possible to hypothesize that the subject was not listening to the associated target keyword. The information is therefore not limited in time to the chosen keyword, but also extended to the associated non-selected keyword in the other stream, and which may not be occurring exactly at the same time. Thus, there are three key moments in the stories, bound by the time limits of the target and the masker keywords, which provide information. For convenience, these key moments will be named *"windows"* in the rest of this chapter. For instance, in Figure 5.2.A, the first target and masker keywords overlap while conversely, the second target and masker keywords are separated by 1 s. Therefore, these key moments, or windows, can have varying duration.

Based on participants' answers, the behavioural stimuli were modelled with three parameters.

Attentional directors

There are two main attentional loci that are not explicitly known and need to be inferred. First of all, onto which story listeners are focusing their attention when there is no behavioural information available. This situation occurs for instance at the very beginning of stories (before the first keyword), at the end of stories (after the last keyword) or between two windows. This first parameter, named here *"attentional director"*, has been modelled in two manners.

1. Dividing the stimulus into three sections modulated by the keywords is a first way of inferring. The durations of these sections are not identical and depend on the temporal positions of the keywords and hence the windows. The cut-out points are halfway between two windows. Figure 5.2.B illustrates how a speech stream can be divided into three sections. In this case, the information contained in the windows is expanded in time.

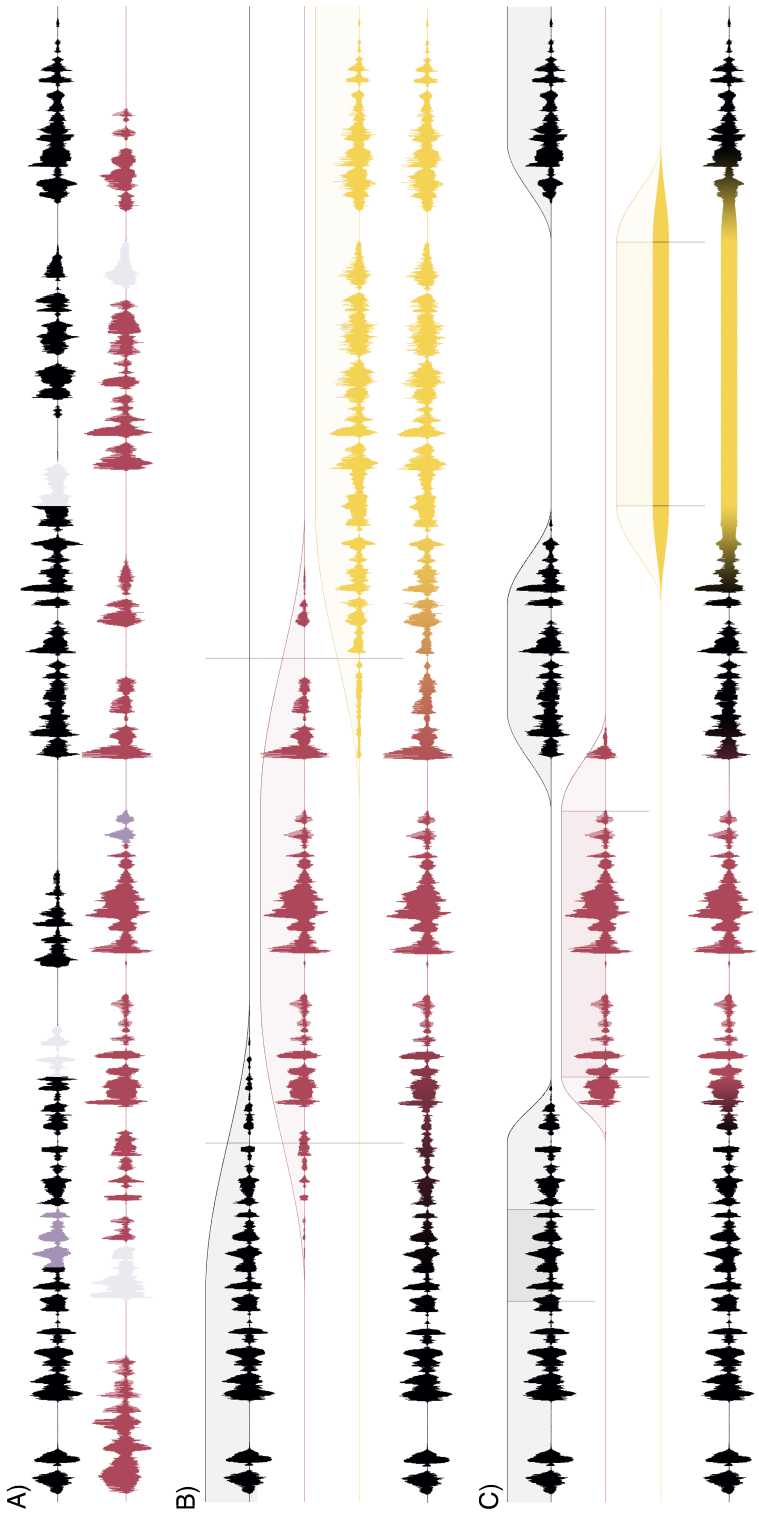

Figure 5.2 : Creation steps of behavioural stimuli. A) represents an example of a trial where the target is in black and the masker is in red. In this example, the participant has answered the first target keyword, the second masker keyword and the third extraneous keyword (highlighted in purple). The remaining keywords are in grey. In B), 3 sections are built according to the participant's answer with a switch attention of 3 s and the extraneous sections filled by the mixture (in yellow). The three sections are then added together. In C), three window directors are built according to the participant's answer with a switch attention of 1 second and the extraneous sections filled by noise. When, there is no information, it is filled with the target. The three windows directors are then added together.

2. A second approach is to assume that the information contained in the window should be limited to windows only. Between the windows, it can be assumed that the participant was listening to either the target or the masker streams. Figure 5.2.C illustrates a situation where the participant listens to the target, even outside the windows, which is the position widely adopted in the literature.

Extraneous keywords

The second attentional locus to infer concerns the extraneous keywords. In addition to being able to choose a keyword that belongs to the target or the masker story, participants can answer a keyword that does not belong to either story: the extraneous keywords. It is likely that the subject was listening more to the mixture of the two streams (illustrated in Figure 5.2.B). Another possibility is that the participant was listening correctly to the target voice (or voice masker) but failed to complete the task, perhaps because they failed to remember the keyword, even though they heard it. Finally, it is possible that the participant was listening neither to the target nor to the masker, and that their attention was wandering off. In these situations, the mixture does not seem the most appropriate acoustic correlate of what the participant is focusing on, and instead, noise (illustrated in Figure 5.2.C) and the envelope of another speech signal were also modelled as control conditions.

Attentional switch

The third parameter is the speed with which participants can switch from one voice to another. The duration of this attentional switch is modelled as the slope of the edges of the time windows. Three values were used: one, two and three seconds.

5.3.4 TRF and Stimulus-reconstruction

The TRF was calculated using the MNE-Python library (Gramfort et al., 2014). The TRF, also called *decoder*, is composed of weights that can be estimated by linear regression for a set of N electrodes at different delays τ. In this experiment, we investigated time delays between -900 ms to 0 ms. The reconstruction of the

speech envelope $\hat{S}(t)$ can be obtained as follows:

$$\hat{S}(t) = \sum_{n=1}^{N} \sum_{\tau} TRF(\tau, n) R(t - \tau, n) \tag{5.1}$$

where R represents the matrix that contains the shifted neural responses of each electrode n at time $t = 1...T$. R and S can be described as:

$$S = [S(0), S(1), S(2), \ldots, S(T)]$$

$$R = \begin{bmatrix} r_1(0) & r_1(1) & \ldots & r_1(\tau) & \ldots & r_1(T) \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & \ldots & r_1(0) & \ldots & r_1(T-\tau) \\ \vdots & \vdots & & & & \vdots \\ r_n(0) & r_n(1) & \ldots & r_n(\tau) & \ldots & r_n(T) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \ldots & r_n(\tau) & \ldots & r_n(T-\tau) \end{bmatrix}$$

.

The matrix R is padded with zeros on the left to ensure causality (see O'Sullivan et al., 2015). A ridge regression can be applied to obtain the weights of the TRF as follows:

$$TRF = (RR^T + \lambda I)^{-1}(RS^T) \tag{5.2}$$

where λ is the regularization parameter, chosen to optimize the stimulus- response reconstruction and I is the identity matrix. The optimal ridge parameter was set to $\lambda = 10^{1/2}$ accordingly to Crosse et al. (2016). TRFs were estimated on a trial-by-trial basis for each subject in each condition. The stimulus-reconstruction of a single trial was predicted in a leave-one-out fashion. To be more precise, each subject had 48 trials per condition. Each trial was reconstructed with the TRF trained on the 47 other trials. The stimulus-reconstruction was evaluated with the Pearson's correlation coefficient between the reconstructed speech envelope and the original speech envelope. This evaluation is usually noted R.

5.3.5 Statistical analyses

All statistics were performed using R (R Core Team, 2017). All the linear mixed models (LMM) were implemented with the *lme4* package (Bates et al., 2014). The models were implemented using a top-down strategy on data (Zuur et al., 2009). The final model is reported with the *lme4* syntax such as Equation 5.3:

$$Score \sim factor_A * factor_B + (factor_A * factor_B \mid subject) \tag{5.3}$$

The full-factorial model is indicated by the fixed effect term $factor_A * factor_B$ and includes main effects and interactions for these two main conditions. The last term of the equation describes an individual random intercept and slope per subject for $factor_A$ and $factor_B$.

For an easier interpretation, the *afex* package (Singmann et al., 2019) was used to compute the statistics of main effects. To do so, the final model was compared to restricted models in which the effect estimated is fixed to 0. Finally, post-hoc analyses were computed a *false discovery rate correction*.

5.4 Results

5.4.1 Stimulus-reconstruction evaluation

Modelling parameters

The three modelling parameters influence was analysed with a linear mixed model (LMM) fitted on the Fisher transformed Pearson's correlation scores (R). Equation 5.4 indicates the final model:

$$R \sim voice * filler + director + (1 \mid subject) \tag{5.4}$$

Similarly to the analysis in Chapter 4, the distance between voices has an effect on participants' performances [$\chi^2(2) = 281.49, p < .001$] but post-doc analyses did not identify any differences between the conditions. Regarding the modelling parameters, the filler for the extraneous word [$\chi^2(4) = 104.12, p < .001$] had an effect on the reconstruction of the behavioural stimuli. Post-hoc analyses showed that when the extraneous word was filled up with the mixture, the stimulus-

reconstruction was the best. Then, the target and the masker streams achieved equally the second best performance. Then came the noise and the other speech stream with the other speech stream that surprisingly performed worse than the noise. The interaction between these two factors $[\chi^2(8) = 26.67, p < .01]$ showed that there was an effect of the distance between voices only when the extraneous keywords were filled with the noise of the other speech stream. Finally, the attentional director $[\chi^2(1) = 9.66, p < .01]$ had also an effect with a better performance when the local windows approach was used over the three sections $[t(20) = -3.13, p < .01]$. Figure 5.3 represents the stimulus-reconstruction for these three factors. It is noteworthy than the attentional switch speed had no effect on participants' performance.

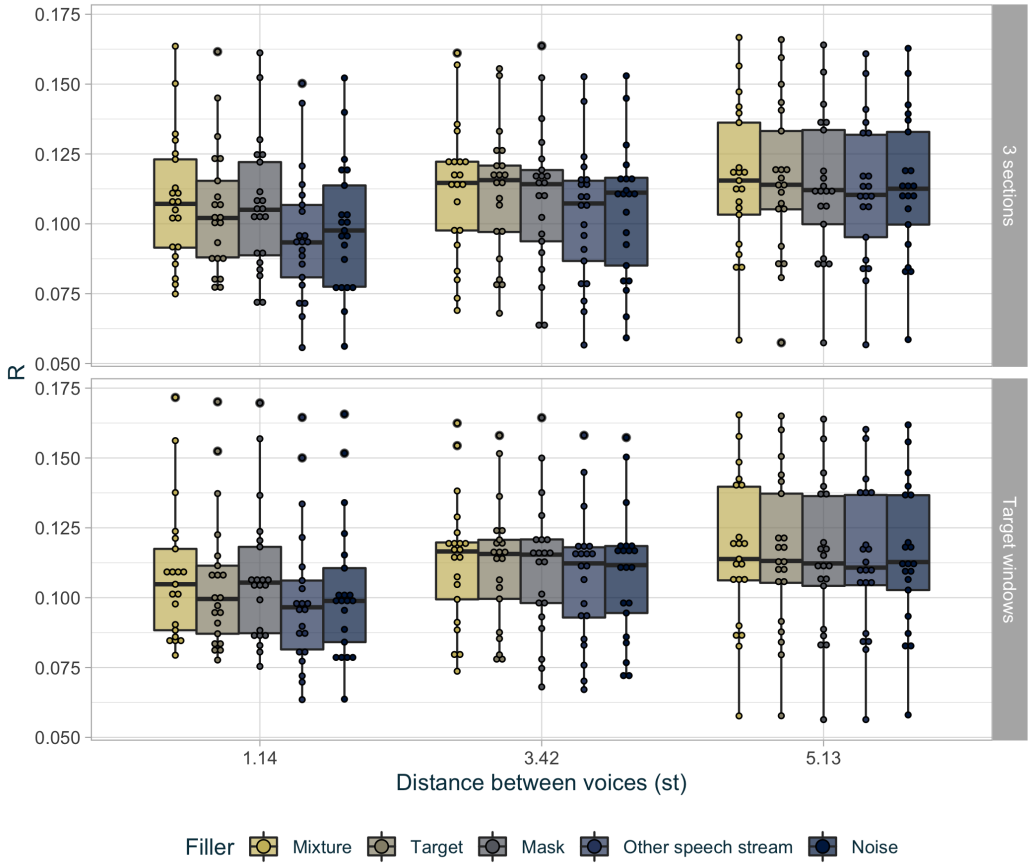

Figure 5.3 : Pearson's correlation for behavioural envelopes. The points represent the scores for every participant in each voice condition for the extraneous keyword (in colour) and the attentional director (top and bottom). The hinges of the boxplot represent the first and the third quartile. The median is represented as a bar in each boxplot. The length of the whiskers is 1.5 interquartile range.

In conclusion, the best stimulus-reconstruction evaluation is modelled when the mixture filled the extraneous keywords with a "local window" process, regardless of the attentional switch duration. Therefore, the analysis in the following section have been carried out with a 2 sec attentional switch.

Original vs. behavioural stimuli

The best modelling reconstruction was compared with the original target reconstruction in this section. The evaluation of these two decoders was analysed with a linear mixed model (LMM) fitted on the Fisher transformed Pearson's correlation scores (R). Equation 5.5 indicates the final model:

$$R \sim voice * decoder + (voice \mid subject) \tag{5.5}$$

Based on the likelihood ratio tests, the distance between the two voice has an effect on participants' scores [$\chi^2(2) = 11.95, p < .01$] as well as the decoder [$\chi^2(1) = 11.1, p < .001$] and the interaction [$\chi^2(2) = 28.85, p < .001$]. Post-hoc analysis showed that participants had higher scores when the best behavioural stimulus was used to reconstruct the speech signal in the most difficult condition [$t(20) = -4.6, p < .001$]. Moreover, there is no difference between stimulus-reconstruction evaluations for the three conditions when the best behavioural stimulus was used to reconstruct the speech signal (see Figure 5.4). The Appendix A shows the comparison for the original target stimulus-reconstruction evaluation compared with all the behavioural stimuli.

5.4.2 Auditory attention decoding

Modelling parameters

The three modelling parameters influence was analysed with a generalized linear mixed model (gLMM) fitted on the binary (correct/incorrect) auditory attention decoding (AAD). Equation 5.6 indicates the final model:

$$AAD \sim voice * director + (1 \mid subject) \tag{5.6}$$

Table 5.1 indicates that results for the the distance between voices were similar to analysis in Chapter 4. Regarding the director, only the interaction with

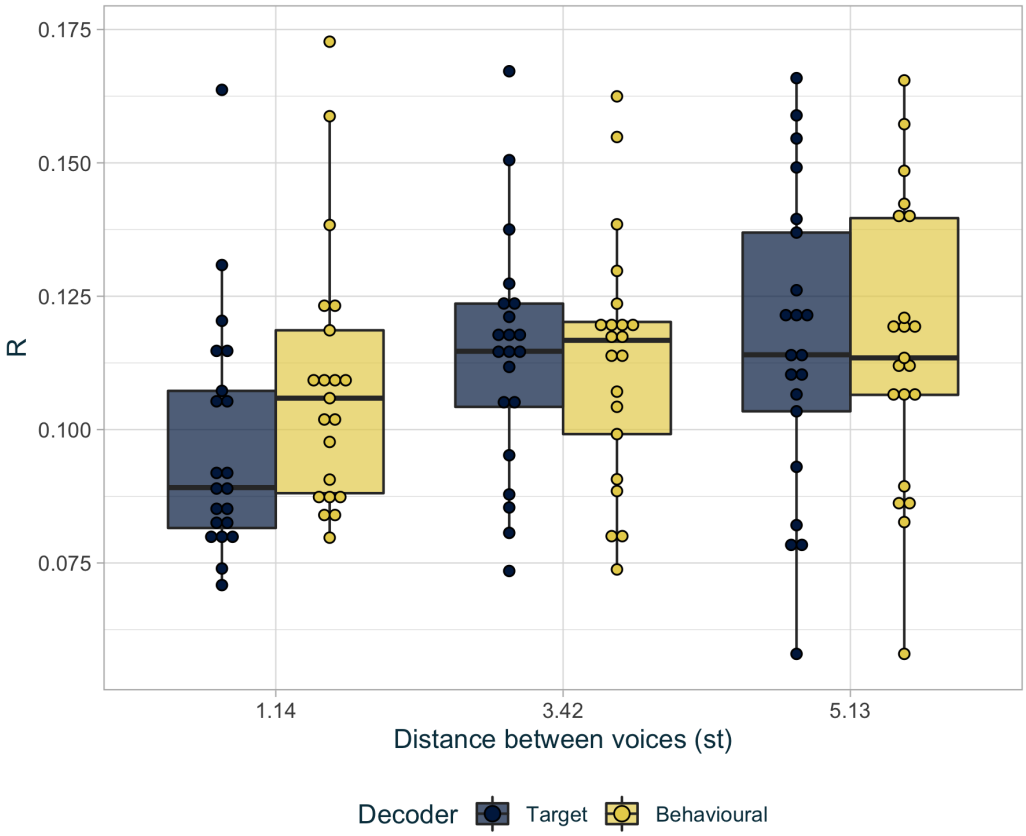

Figure 5.4 : Pearson's correlation for the original target stimuli (in dark blue) and for the best behavioural stimuli (in light yellow).

the voice was significant (see Figure 5.5). In conclusion, the three modelling parameters don't modulate AAD performances. Therefore, the analysis in the following section have been carried out with the same parameters than the best stimulus-reconstruction behavioural stimulus.

Table 5.1 gLMM coefficients for the distance between the two voices (centered on 0), the director and their interaction as fixed factors fitted on correct/incorrect data.

Fixed effects	Coefficients		Statistics	
	β	SE	z	p
Intercept	0.87	0.07	11.04	$< .001$
Voice	0.6	0.02	33.53	$< .001$
Director	0.001	0.01	0.15	.88
Voice × Director	-0.04	0.02	-2.22	$< .05$

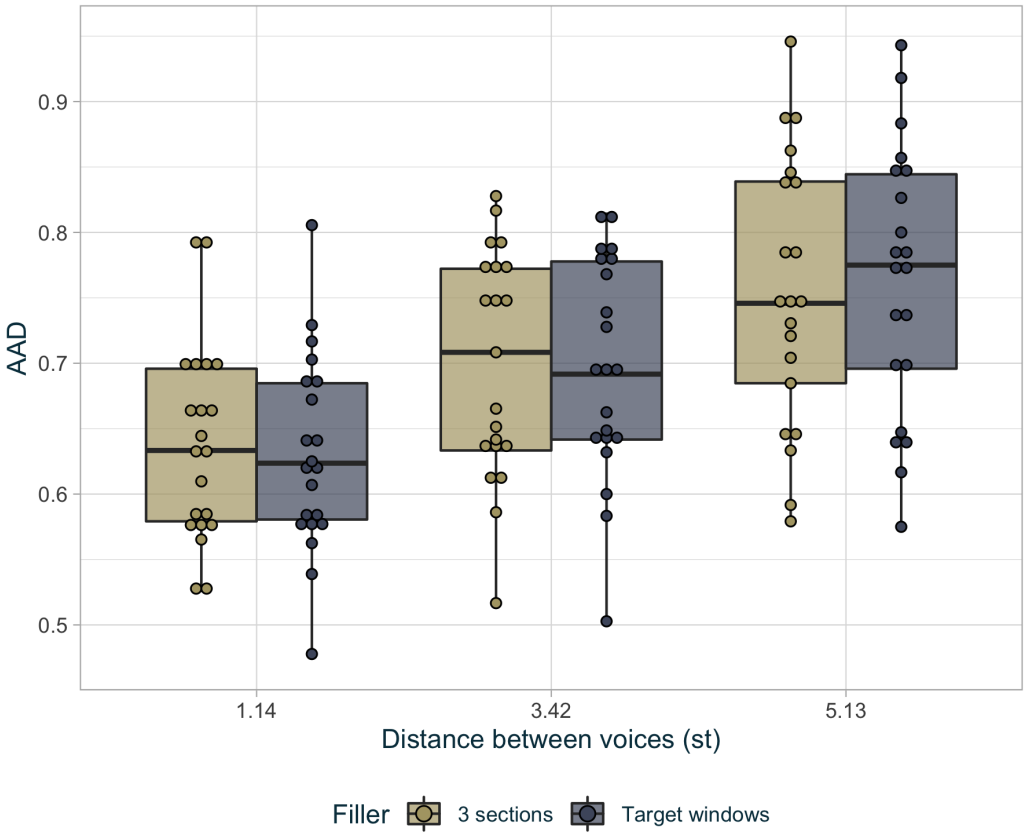

Figure 5.5 : AAD for behavioural envelopes. The points represent the scores for every participant in each voice condition for the attentional director.

Original vs. behavioural stimuli

The best modelling AAD was compared with the original target AAD in this section. The evaluation of these two decoders was analysed with a generalized linear mixed model (gLMM) fitted on the binary (correct/incorrect) auditory attention decoding (AAD. Equation 5.7 indicates the final model:

$$AAD \sim voice + (voice \mid subject) \tag{5.7}$$

Results show there is no decoder effect. The Appendix A shows the comparison for the original target AAD compared with all the behavioural stimuli.

5.5 Discussion and conclusion

The goal of this chapter was to assess if information regarding the attentional focus, inferred from participants' answer to the *Long-SWoRD* test, could provide a better reconstruction evaluation. This is the case only in challenging listening situations, where the participants' attention is less likely to remain focused entirely on the target talker. On the other hand, no improvement was observed in easy conditions. One explanation could be the reduced number of errors. Behavioural stimuli are based on the participants' answers. Thus in the absence of errors (or in the presence of a faultless trial), the behavioural stimulus will be identical to the original target. It is not surprising that under easy conditions, there are minimal differences between the behavioural stimuli and the original target. However, although these differences are not significant, the reconstructions performed with the original target seem to outperform those with behavioural stimuli. Thus, in situations where the two competing voices are clearly distinct and easily separated perceptually, the assumption that listeners are able to stay focused on the target could be reasonable.

Several parameters were used to model the behavioural stimuli. Extraneous word filling seems to be the most important factor, while the attentional director and the duration of attentional switches do not seem to really influence the results. If the extraneous word is replaced by the mixture, the target or even the masker stream, the performance of the reconstruction is improved (in the challenging condition).

Regarding the AAD, no difference was found between the behavioural stimuli and the original target. In addition, the modelling parameters had no influence on the AAD. However, the differences, although not significant, seemed to follow the same pattern as in the reconstruction of stimuli: improvement under challenging conditions. However, the computational power and the fact that we used a diotic presentation of the stimuli might explain the lack of improvement. AAD is not always above chance, especially in a challenging condition, as the results in Chapter 4 point out. If Mirkovic et al. (2015) emphasize that it takes at least 15 one-minute trials to observe performance above chance, a slightly different configuration has been used with the Long-SWoRD test: 47 trials of 11-18 s leading up to 11-12 minutes for training the data. With regard to the presentation of diotic stimuli, few TRF studies have been published, most using dichotic stimuli. Fiedler et al. (2019) stands out as an exception and their AAD varies between 55 and 62%, which is low compared to a AAD with dichotic stimuli. As such, although our AAD appears to be low, it is consistent with the available literature.

The *Long-SWoRD* test provides experimenters with a means to behaviourally track fluctuations in auditory selective attention. Teoh and Lalor (2019) improved auditory attention decoding accuracy by incorporating spatial attentional focus. Results presented in this chapter show that this enhancement can also be achieved in challenging situations where attention is modulated by voice cues such as F0 and VTL. By monitoring the attentional focus, it is possible to obtain a better reconstruction of the real attended speech and therefore a better cortical representation.

Chapter 6

General discussion, conclusion and future perspectives

A considerable amount of research has brought together the accumulated understanding of the cocktail party effect. While behavioural and neurophysiological approaches seek to better understand how a listener solves the auditory scene analysis, it appears that there is a lack of harmonization of tasks and stimuli (Chapter 1 for more information). With the idea to overcome this shortcoming, an intelligibility task has been designated - the *Long-SWoRD* test - and presented in Chapter 2. Participants' results, behavioural as well as neurophysiological, were respectively reported in Chapter 3 and Chapter 4. Finally, Chapter 5 explores the combination of behavioural and neurophysiological measures for modelling fluctuations in auditory selective attention. In this present Chapter, the main results and limitations are discussed, respectively in Section 6.1 and Section 6.2 while future perspectives are presented in Section 6.3. Finally, the conclusions are briefly summarized in the Section 6.4.

6.1 The Long-SWoRD test

6.1.1 Segregation, cognitive resources and semantic context

The first goal of this manuscript was to document the performance of normal-hearing listeners in this task in situations where the perceptual separability of the competing voices ranges from easy to hard using a combination of voice and spatial cues. First and foremost, it is important to stress the coherence of the behavioural outputs across experiments 1 and 2 from Chapter 3 and experiment 3 from Chapter 4. Second, in general, participants' behavioural answers reflect the same pattern as performances measured with popular tests such as CRM (Bolia et al., 2000): the more dissimilar target and masker voices are, the greater the performance. In addition, a ceiling effect was observed for participants' performances when the difficulty of the Long-SWoRD was not high enough.

The second purpose was to bridge the gap between behavioural and neurophysiological studies and to observe if the same results were obtained using the two different approaches. Hence, with regard to neurophysiological studies, Temporal Response Function (TRF) analyses were conducted for their advantage in analysing long duration stimuli. Once again, the pattern of observed performances is consistent with the literature: the auditory attention decoding (AAD), an indication that the representation of the speaker to whom our selective attention is directed, improves as the task gets easier. In addition, the target speech reconstruction and hence its cortical representation, is better than the masker speech reconstruction. This difference between the two representations increases as target and masker voices become dissimilar but seems to level off at a certain point. This plateau is also reflected in the behavioural measurement of errors. As discussed in Chapter 4, it is possible that, once speakers' representations are optimal, the amount of cognitive resources remaining determines behavioural performance and the AAD. This hypothesis could notably be supported by studies on energetic and informational masking. Informational masking seems to arise due to competition for resources at central cortical auditory processing levels: the more similar target and masker voices are, the higher the activation is (for a review, see Scott & McGettigan, 2013).

In contrast to other behavioural studies, in the Long-SWoRD test, participants can benefit from schema-based cues described by Bregman (1994), such as lin-

guistic knowledge, to segregate two speakers. The semantic context contribution can increase performances as early modelling suggests (Boothroyd & Nittrouer, 1988). Unsurprisingly, Long-SWoRD performances are better than the literature (although once again, they follow the same pattern). The implementation of a semantic context measure, however, indicates that this semantic contribution is limited to challenging conditions, which is in line, once again, with results of studies in the field of energetic and informational masking (for a review, see Mattys et al., 2012). In addition, the presence of N400 provides an understanding of how semantic contexts are exploited by listeners. Interestingly, and in contrast to Broderick et al.'s study (2018), it appears that participants rely on semantic cues from target and masker streams to perform the task. However, it should be noted that the latter assumption implies that, to some extent, participants share their attention on both talkers. According to Shinn-Cunningham et al. (2017), listeners, instructed in advance to report back two messages, are able to divide attention between multiple simultaneous streams. However, the ability to share attention seems to be limited to short messages and not too demanding tasks. It is therefore unlikely that participants were able to maintain a constantly divided attention on the two streams. It is possible, though, that listeners may share their attention repeatedly for a few seconds to develop a vague idea of the semantic context of the masker story.

6.1.2 From switching attention to the build-up effect

From the combination of Long-SWoRD and TRF, two new perspectives have emerged. Firstly, the Long-SWoRD test was designed to provide experimenters with a means to behaviourally track fluctuations in auditory selective attention. The information regarding the actual — as opposed to, assumed — attentional focus can be used advantageously during model training, to enhance subsequent (test phase) accuracy of auditory stimulus-reconstruction based on EEG signals. This is the case only in challenging listening situations, where the participants' attention is less likely to remain focused entirely on the target talker.

Secondly, the analysis of the decomposition of the mixture into two streams over time with TRF (*build-up effect*) seems to support previous behavioural studies. Our results show that the build-up of the two streams would be approximately 5-6 seconds, which supports behavioural studies suggesting that 2 to 10 seconds is sufficient (Best et al., 2018; Bregman, 1994; B. C. J. Moore & Gockel, 2012).

To our knowledge, this is the first time that neurophysiological analyses have confirmed this duration in humans. Moreover, usually the build-up effect is studied with simple stimuli such as syllables or pure tones, while our study documents the build-up of more complex stimuli such as short stories.

As discussed in Chapter 4, there isn't a clear segregation of voices in the very challenging condition and the build-up is never really complete. According to B. C. J. Moore and Gockel (2012) the target and the masker build up over time but may be partially or completely reset by a sudden change in the properties of the speeches signal or by switches in attention. A recent MEG study (Billig, Davis, & Carlyon, 2018) investigated the segregation and grouping of sequences of pure high (H) and low (L) tones presented in a repeating pattern HLH. They showed that the comparison between segregation and grouping of two streams is reflected from 66 to 138 ms after the onset of the L tone, which the authors interpreted as a P1 marker. In the studies presented in this thesis, a difference between the P1 markers from the target and the masker streams is precisely present in the conditions where a build-up effect is observed. On the contrary, this difference of P1 is absent in the very challenging condition, where the build-up does not fully occurs. The analysis of the tendency to separate two streams over time seems therefore seems consistent with the literature.

6.2 Limitations

The Long-SWoRD suffers from two main shortcomings. The first limitation has already been mentioned in Chapter 3 and concerns the approach with which masker voices were synthesized. This process was essential to quantify the distance between two voices, but it is coupled with the sacrifice of ecological validity (Lavan, Knight, & McGettigan, 2019). Voice discrimination is not two-dimensional (F0 and VTL) but highly multi-dimensional. Thus, unlike in real life, our participants could not benefit from cues such as speaking style or prosody to discriminate two speakers. Furthermore, a recent review (Myers, Lense, & Gordon, 2019) argues that prosody has an inherent role in speech perception and, hence, in neural entrainment. Since prosodic cues are represented in the amplitude envelope, neural tracking inherently capture a response to prosody to some degree. For instance, Ding et al. (2017) showed that neural entrainment strengthens with prosodic cues. In our experiments, participants can only benefit from within-

speakers differences in prosody and not from across-speakers differences in prosody. As such, the speech tracking was perhaps restricted.

The second limitation refers to the decision-making aspect of the Long-SWoRD and a potential uncontrolled memory effect. Participants have to make a decision about the presence of keywords in the target story, retroactively after listening to the stimuli which can last up to 18 seconds. The task asked of the participants requires several steps: (1) listening to the right story, (2) encoding the information, (3) storing the information,(4) retrieving the information and (5) reaching a decision. The answer given by the participant is therefore influenced by intelligibility, working memory and the decision-making aspect. However, the last two are not controlled and are not accounted for in the analyses which can affect the validity of behavioural data. For instance, although behavioural and neurophysiological performance appear to follow the same response pattern, the correlation between these two measures remains weak at an individual level at an equal level of difficulty (see Chapter 4). Thus, good tracking speech score in a particular participant does not necessarily lead to good performance on the Long-SWoRD. However, top-down inputs such as working memory can influence neural tracking as Hjortkjaer et al. (2018) pointed out. It is therefore likely that cognitive resources such as working memory could modulate the link between intelligibility and speech tracking.

6.3 Future perspectives

6.3.1 Upgrading the Long-SWoRD test

Several modifications to the Long-SWoRD protocol and analyses could directly address the last limitation aforementioned. In a review, Wilsch and Obleser (2016) highlighted the different roles of neural oscillations in the theta ($\sim 4-8$ Hz), alpha ($\sim 8-13$ Hz), and gamma ($\sim 30-200$ Hz) frequency range. While theta power has been repeatedly found to increase with working memory load, gamma power presumably reflects active maintenance of working memory representations. Furthermore, the authors particularly emphasize how alpha power can reflect memory load in auditory working memory. The studies presented in this thesis are limited to the theta frequency range. An alpha-wave analysis could provide more information on working memory activity. Similarly, getting participants

to complete a span task could also inform about individual working memory capacities.

Finally, even when obtaining working memory information, the decision-making aspect of the task remains a black box. Participants are confronted with several options when they have to choose the keywords. It is very likely that these options will influence their choices. For example, as mentioned in Chapter 4, participants may use the semantic context of the masker to help them succeed in the task: by knowing the masker words, participants limit their mistakes. Thus, the resolution of this task does not only rely on their ability to select the target keyword, but also on their ability to not select the masker keyword. Instead of asking subjects to choose a keyword, they could be asked to rank the different keywords, depending on the certainty that participants have of the keyword presence in the target story. If participants rely on the masker semantic context, their ranking is likely to be (1) target keyword, (2) extraneous keyword, and (3) masker keyword. In this way, the masker keyword should be ranked last since participants would be assured that it is not part of the target stream. Finally, additional metrics such as the response time or the number of times participants corrected their answer could also provide informations about the difficulty to reach a decision.

6.3.2 Towards the irrelevant sound effect

As mentioned in Chapter 1, the irrelevant sound effect (ISE) refers to the degradation of a working memory task when speech sound is, verbally or visually, presented. One of the assumptions proposed for the ISE is the presence of fluctuations in the spectro-temporal fine structure of the masker stream (Beaman et al., 2007; Ellermeier & Zimmer, 2014).

Throughout this work, particular emphasis was placed on challenging listening situations since there was no apparent effect of acoustic and semantic cues when the target and masker talkers were too dissimilar or presented dichotically. As such, the Long-SWoRD test appears to be a simple test of working memory under easy listening conditions. Ding, Chatterjee, and Simon (2014) showed that degrading the fine structure of a speech streams weakens the neural entrainment. Thus, by degrading the spectro-temporal fine structure of the masker and by measuring what level of speech tracking these signals induce, the Long-SWoRD test could provide a means to test the fine structure of the distractor stream effect.

The results could then be used to design new methods for systematic evaluation of the ISE.

6.3.3 Practical and clinical applications

The enhancement of hearing aids is essential to enable hearing impaired listeners to focus more easily on a speaker in noise. The advantage of parametric voice manipulation, as introduced in the studies presented in this thesis, is that the listening difficulty can be controlled. By generating extremely challenging conditions, it is possible to approach listening situations similar to that experienced by people with hearing loss and cochlear implants (CI) on a daily basis. For instance, CI users do not seem to efficiently benefit from voice cues, such as F0 and VTL, to discriminate two speech streams (El Boghdady, Gaudrain, & Başkent, 2019; Gaudrain & Başkent, 2018). This was also the case for the participants of the studies presented in this thesis, under the conditions of challenging listening.

One of the major challenges in neural tracking studies is to identify, based on brain activity, the speaker that the participant is listening to in a cocktail party situation. However, as shown in the Chapter 4, attentional decoding performance is weak in challenging situations. These results highlight the difficulty of implementing efficient algorithms in hearing aid technologies for users in cocktail party situations.

The results of the third study, presented in Chapter 4, suggest that, in situations of adverse listening, semantic cues play a crucial role in the separation of two speech streams. Thus, it will probably be essential to incorporate measures of this semantic information in future hearing aids technology to offer people with hearing loss an effective solution to achieve easy and comfortable listening in noise.

6.4 Conclusion

In these studies, we have shown that:

1. While parametrically manipulated voice cues affect the perception of competing talker, semantic cues also play a quantifiable role, provided enough semantic context is present.

2. AAD does not efficiently reflect speech segregation in challenging situations where the target-masker voice difference is not very salient. This raises questions of the applicability of AAD as a support to hearing impaired listeners in challenging cocktail party situations.

3. Across different levels of difficulty, the TRFs reflect that different processes are used: for clearly separable speakers, sensory cues are very obvious and both stream can be separately processed; at intermediate difficulty, sensory cues dominate, and are complemented with linguistic cues; at the highest difficulty, only late linguistic processing seems to be differentiate target and masker.

4. Combining behavioural data with EEG in the training phase of TRFs improves the quality of reconstruction in the most challenging condition.

References

Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, *9*(1), 1–12. doi: 10.1038/s41598-018-37359-z

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., & Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage*, *124*, 906–917. doi: 10.1016/j.neuroimage.2015.09.048

Akram, S., Simon, J. Z., & Babadi, B. (2017). Dynamic Estimation of the Auditory Temporal Response Function from MEG in Competing-Speaker Environments. *IEEE transactions on bio-medical engineering*, *64*(8), 1896–1905. doi: 10.1109/TBME.2016.2628884

Alain, C., Reinke, K., He, Y., Wang, C., & Lobaugh, N. (2005). Hearing Two Things at Once: Neurophysiological Indices of Speech Segregation and Identification. *Journal of Cognitive Neuroscience*, *17*(5), 811–818. doi: 10.1162/0898929053747621

Alain, C., Reinke, K., McDonald, K. L., Chau, W., Tam, F., Pacurar, A., & Graham, S. (2005). Left thalamo-cortical network implicated in successful speech separation and identification. *NeuroImage*, *26*(2), 592–599. doi: 10.1016/j.neuroimage.2005.02.006

Başkent, D., & Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *The Journal of the Acoustical Society of America*, *139*(3), EL51-EL56. doi: 10.1121/1.4942628

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beaman, C. P., Bridges, A. M., & Scott, S. K. (2007). From Dichotic Listening to the Irrelevant Sound Effect: A Behavioural and Neuroimaging Analysis of the Processing of Unattended Speech. *Cortex*, *43*(1), 124–134. doi: 10.1016/S0010-9452(08)70450-7

Bednar, A., & Lalor, E. C. (2020). Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. *NeuroImage*, *205*, 116283. doi: 10.1016/j.neuroimage.2019.116283

Best, V., Keidser, G., Buchholz, J. M., & Freeston, K. (2016). Development and preliminary evaluation of a new test of ongoing speech comprehension. *International journal of audiology*, *55*(1), 45–52. doi: 10.3109/14992027.2015.1055835

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, *105*(35), 13174–13178. doi: 10.1073/pnas.0803718105

Best, V., Streeter, T., Roverud, E., Mason, C. R., & Kidd, G. (2016). A Flexible Question-and-Answer Task for Measuring Speech Understanding. *Trends in Hearing*, *20*, 233121651667870. doi: 10.1177/2331216516678706

Best, V., Swaminathan, J., Kopčo, N., Roverud, E., & Shinn-Cunningham, B. (2018). A "buildup" of speech intelligibility in listeners with normal hearing and hearing loss. *Trends in Hearing*, *22*, 233121651880751. doi: 10.1177/2331216518807519

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2017). Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(5), 402–412. doi: 10.1109/TNSRE.2016.2571900

References

Bilger R. C., Nuetzel J. M., Rabinowitz W. M., & Rzeczkowski C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, *27*(1), 32–48. doi: 10.1044/jshr.2701.32

Billig, A. J., Davis, M. H., & Carlyon, R. P. (2018). Neural Decoding of Bistable Sounds Reveals an Effect of Intention on Perceptual Organization. *Journal of Neuroscience*, *38*(11), 2844–2853. doi: 10.1523/JNEUROSCI.3022-17.2018

Billig, A. J., Davis, M. H., Deeks, J. M., Monstrey, J., & Carlyon, R. P. (2013). Lexical Influences on Auditory Streaming. *Current Biology*, *23*(16), 1585–1589. doi: 10.1016/j.cub.2013.06.042

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature reviews. Neuroscience*, *14*(10), 693–707. doi: 10.1038/nrn3565

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066. doi: 10.1121/1.428288

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, *84*(1), 101–114. doi: 10.1121/1.396976

Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 380–387. doi: 10.1037/0096-1523.4.3.380

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound.* MIT Press.

Broadbent, D. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, *47*(3), 191–196. doi: 10.1037/h0054182

Broadbent, D. (1958). *Perception and communication.* Elmsford: Pergamon Press. doi: 10.1037/10037-000

Brocolini, L., Parizet, E., & Chevret, P. (2016). Effect of masking noise on cognitive performance and annoyance in open plan offices. *Applied Acoustics*, *114*, 44–55. doi: 10.1016/j.apacoust.2016.07.012

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. doi: 10.1016/j.cub.2018.10.042

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809. doi: 10.1016/j.cub.2018.01.080

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *Journal of Neuroscience*, *39*(38), 7564–7575. doi: 10.1523/JNEUROSCI.0584-19.2019

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487. doi: 10.3758/s13414-015-0882-9

Brouwer, S. (2017). Masking release effects of a standard and a regional linguistic variety. *The Journal of the Acoustical Society of America*, *142*(2), EL237-EL243. doi: 10.1121/1.4998607

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*, *131*(2), 1449–1464. doi: 10.1121/1.3675943

Brungart, D. S. (2001). Evaluation of speech intelligibility with the coordinate response measure. *The Journal of the Acoustical Society of America*, *109*(5), 2276–2279. doi: 10.1121/1.1357812

Brungart, D. S., & Simpson, B. D. (2007). Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task. *The Journal of the Acoustical*

Society of America, *122*(3), 1724–1734. doi: 10.1121/1.2756797

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. doi: 10.1121/1.1408946

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., & Bradlow, A. R. (2013). Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. *American Journal of Audiology*, *22*(1), 157–164. doi: 10.1044/1059-0889(2013/12-0072)

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America*, *128*(2), 860–869. doi: 10.1121/1.3458857

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 115–127. doi: 10.1037/0096-1523.27.1.115

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), 975–979. doi: 10.1121/1.1907229

Cherry, E. C., & Taylor, W. K. (1954). Some Further Experiments upon the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *26*(4), 554–559. doi: 10.1121/1.1907373

Chiba, T., & Kajiyama, M. (1941). *The vowel: Its nature and structure*. Tokyo: Tokyo-Kaiseikan.

Chomsky, N. (2002). *Syntactic Structures*. Walter de Gruyter.

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., ... Smalt, C. J. (2019). Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods. *Scientific Reports*, *9*(1), 1–10. doi: 10.1038/s41598-019-47795-0

Clarke, J., Gaudrain, E., Chatterjee, M., & Başkent, D. (2014). T'ain't the way you say it, it's what you say – Perceptual continuity of voice and top–down restoration of speech. *Hearing Research*, *315*, 80–87. doi: 10.1016/j.heares.2014.07.002

Clopper, C. G., Pisoni, D. B., & Tierney, A. T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology*, *17*(5), 331–349.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*(2), 331–335. doi: 10.3758/BF03196169

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, *96*(2), 341–370.

Cowan, N., & Wood, N. L. (1997). Constraints on Awareness, Attention, Processing, and Memory: Some Recent Investigations with Ignored Speech. *Consciousness and Cognition*, *6*(2), 182–203. doi: 10.1006/ccog.1997.0300

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. doi: 10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. doi: 10.3389/fnhum.2016.00604

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 643–656. doi: 10.1037/0096-1523.30.4.643

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913–2922. doi: 10.1121/1.1616924

de Cheveigné, A. (1999). Waveform interactions and the segregation of concurrent vowels. *The Journal of the Acoustical Society of America*, *106*(5), 2959–2972. doi: 10.1121/1.428115

de Cheveigné, A., Wong, D. D. E., Di Liberto, G. M., Hjortkjær, J., Slaney, M., & Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, *172*, 206–216. doi: 10.1016/j.neuroimage.2018.01.033

de Taillez, T., Kollmeier, B., & Meyer, B. T. (2018). Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*. doi: 10.1111/ejn.13790

Decruy, L., Vanthornhout, J., & Francart, T. (2019). Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. *Journal of Neurophysiology*, *122*(2), 601–615. doi: 10.1152/jn.00687.2018

Dekerle, M., Boulenger, V., Hoen, M., & Meunier, F. (2014). Multi-talker background and semantic priming effect. *Frontiers in Human Neuroscience*, *8*. doi: 10.3389/fnhum.2014.00878

Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech. *eNeuro*, *5*(2). doi: 10.1523/ENEURO.0084-18.2018

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, *25*(19), 2457–2465. doi: 10.1016/j.cub.2015.08.030

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46. doi: 10.1016/j.neuroimage.2013.10.054

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing Neural Entrainment to Hierarchical Linguistic Units using Electroencephalography (EEG). *Frontiers in Human Neuroscience*, *11*. doi: 10.3389/fnhum.2017.00481

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. doi: 10.1073/pnas.1205381109

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. doi: 10.1152/jn.00297.2011

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*, *116*(20), 10113–10121. doi: 10.1073/pnas.1816414116

Drennan, D. P., & Lalor, E. C. (2019). Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity. *eNeuro*, *6*(3). doi: 10.1523/ENEURO.0082-19.2019

Egan, J. P., Carterette, E. C., & Thwing, E. J. (1954). Some factors affecting multi-channel listening. *The Journal of the Acoustical Society of America*, *26*(5), 774–782. doi: 10.1121/1.1907416

El Boghdady, N., Gaudrain, E., & Başkent, D. (2019). Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users? *The Journal of the Acoustical Society of America*, *145*(1), 417–439. doi: 10.1121/1.5087693

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron*, *61*(2), 317–329. doi: 10.1016/j.neuron.2008.12.005

Ellermeier, W., & Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science and Technology*, *35*(1), 10–16. doi: 10.1250/ast.35.10

Enders, G., Enders, J., & Liber, I. (2015). *Le charme discret de l'intestin: tout sur un organe mal aimé*. Paris: Éditions de Noyelles.

Enders, G., Enders, J., & Shaw, D. (2015). *Gut: The inside story of our body's most under-rated organ*. Scribe Publications.

Enders, G., Monceau, J., & Liber, I. (2016). *Le Charme discret de l'intestin: Livre audio 1 CD MP3*. Paris: Audiolib.[Audiobook].

Ericson, M. A., & McKinley, R. L. (2001). *The intelligibility of multiple talkers separated spatially in noise* (Tech. Rep. No. AFRL-HE-WP-SR-2001-0009). Air Force Research Laboratory Wright-Patterson AFB OH Human Effectiveness Directorate.

Fant, G. (1970). *Acoustic theory of speech production.* Walter de Gruyter.

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., & Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of Neural Engineering*, *14*(3), 036020. doi: 10.1088/1741-2552/aa66dd

Fiedler, L., Wöstmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage*, *186*, 33–42. doi: 10.1016/j.neuroimage.2018.10.057

Flego, S. (2018). Estimating vocal tract length by minimizing non-uniformity of cross-sectional area. *Proceedings of Meetings on Acoustics*, *35*(1), 060003. doi: 10.1121/2.0001000

Fontan, L., Tardieu, J., Gaillard, P., Woisard, V., & Ruiz, R. (2015). Relationship Between Speech Intelligibility and Speech Comprehension in Babble Noise. *Journal of Speech, Language, and Hearing Research*, *58*(3), 977–986. doi: 10.1044/2015_JSLHR-H-13-0335

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *109*(5), 2112–2122. doi: 10.1121/1.1354984

Fuglsang, S. A., Dau, T., & Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage*, *156*, 435–444. doi: 10.1016/j.neuroimage.2017.04.026

Gaudrain, E., & Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *The Journal of the Acoustical Society of America*, *137*(3), 1298–1308. doi: 10.1121/1.4908235

Gaudrain, E., & Başkent, D. (2018). Discrimination of voice pitch and vocal-tract length in cochlear implant users. *Ear and Hearing*, *39*(2), 226–237. doi: 10.1097/AUD.0000000000000480

Gaudrain, E., & Crouzet, O. (2019). *word2vec model trained on lemmatized French Wikipedia 2018.* Zenodo. doi: 10.5281/zenodo.3241447

Gautreau, A., Hoen, M., & Meunier, F. (2015). Lexical decision task on French target words: Effect of listeners' knowledge of the babble-language. *Speech Communication*, *69*, 9–16. doi: 10.1016/j.specom.2015.02.004

Geirnaert, S., Francart, T., & Bertrand, A. (2020). An Interpretable Performance Metric for Auditory Attention Decoding Algorithms in a Context of Neuro-Steered Gain Control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(1), 307–317. doi: 10.1109/TNSRE.2019.2952724

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., & Collet, L. (2000). Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency. *The Journal of the Acoustical Society of America*, *108*(1), 263–271. doi: 10.1121/1.429462

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., & Oxenham, A. J. (2005). Neuromagnetic Correlates of Streaming in Human Auditory Cortex. *Journal of Neuroscience*, *25*(22), 5382–5388. doi: 10.1523/JNEUROSCI.0347-05.2005

Hafter, E. R., Xia, J., & Kalluri, S. (2013). A naturalistic approach to the cocktail party problem. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, & H. E. Gockel (Eds.), *Basic Aspects of Hearing* (pp. 527–534). Springer New York.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*, *11*(2), 79–87. doi: 10.3109/01050398209076203

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann,

F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Hausfeld, L., Riecke, L., Valente, G., & Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage*, *181*, 617–626. doi: 10.1016/j.neuroimage.2018.07.052

Helfer, K. S., & Freyman, R. L. (2009). Lexical and indexical cues in masking by competing speech. *The Journal of the Acoustical Society of America*, *125*(1), 447–456. doi: 10.1121/1.3035837

Hjortkjaer, J., Märcher-Rørsted, J., Fuglsang, S. A., & Dau, T. (2018). Cortical oscillations and entrainment in speech processing during working memory load. *European Journal of Neuroscience*. doi: 10.1111/ejn.13855

Hoen, M., Meunier, F., Grataloup, C.-L., Pellegrino, F., Grimault, N., Perrin, F., ... Collet, L. (2007). Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. *Speech Communication*, *49*(12), 905–916. doi: 10.1016/j.specom.2007.05.008

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science*, *29*(10), 1575–1583. doi: 10.1177/0956797618779083

Huet, M.-P., Micheyl, C., Gaudrain, E., & Parizet, E. (2018). Who are you listening to? Towards a dynamic measure of auditory attention to speech-on-speech. *Interspeech 2018*, 2272–2275.

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of speech, language, and hearing research : JSLHR*, *51*(3), 562–573. doi: 10.1044/1092-4388(2008/040)

IEEE Audio and Electroacoustics Group. (1969). *IEEE recommended practice for speech quality measurements.* New York, NY: Institute of Electrical and Electronics Engineers.

Ives, D. T., Vestergaard, M. D., Kistler, D. J., & Patterson, R. D. (2010). Location and acoustic scale cues in concurrent speech recognition. *The Journal of the Acoustical Society of America*, *127*(6), 3729–3737. doi: 10.1121/1.3377051

Iyer, N., Brungart, D. S., & Simpson, B. D. (2010). Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task. *The Journal of the Acoustical Society of America*, *128*(5), 2998–3010. doi: 10.1121/1.3479547

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. doi: 10.1177/0956797613482467

Jones, D., Alford, D., Bridges, A., Tremblay, S., & Macken, B. (1999). Organizational factors in selective attention: The interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 464–473. doi: 10.1037/0278-7393.25.2.464

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351. doi: 10.1121/1.381436

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*(3-4), 187–207. doi: 10.1016/S0167-6393(98)00085-5

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54*(sup2), 3–16. doi: 10.3109/14992027.2015.1020971

Kong, Y.-Y., Mullangi, A., & Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research*, *316*, 73–81. doi: 10.1016/j.heares.2014.07.009

Kostallari, K., Parizet, E., Chevret, P., Amato, J.-N., & Galy, E. (2020). Irrelevant speech effect in open plan offices: Comparison of two models explaining the decrease in performance by speech intelligibility and attempt to reduce interindividual differences of the mental workload by task customisation. *Applied Acoustics*, *161*, 107180. doi: 10.1016/j.apacoust .2019.107180

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59. doi: 10.1080/ 23273798.2015.1102299

Lafon, J.-C. (1964). *Le Test phonétique et la mesure de l'audition*. Paris; Eindhoven: Dunod ; Ed. Centrex.

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189–193. doi: 10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *Journal of Neurophysiology*, *102*(1), 349–359. doi: 10.1152/jn.90896.2008

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 2404. doi: 10.1038/ s41467-019-10295-w

Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., & Francart, T. (2019). Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations. *bioRxiv*, 471367. doi: 10.1101/471367

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010. doi: 10.1016/j.neuron .2007.06.004

Markides, A. (1978). Whole-word scoring versus phoneme scoring in speech audiometry. *British Journal of Audiology*, *12*(2), 40–46. doi: 10.3109/03005367809078852

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

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

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. doi: 10.1038/nature11020

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex. *Journal of Neurophysiology*, *102*(6), 3329–3339. doi: 10.1152/jn.91128.2008

Micheyl, C., Kreft, H., Shamma, S., & Oxenham, A. J. (2013). Temporal coherence versus harmonicity in auditory stream formation. *The Journal of the Acoustical Society of America*, *133*(3), EL188-EL194. doi: 10.1121/1.4789866

Micheyl, C., & Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing Research*, *266*(1), 36–51. doi: 10.1016/j.heares.2009.09.012

Middlebrooks, J. C. (2017). Spatial stream segregation. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party* (pp. 137–168). Cham: Springer International Publishing. doi: 10.1007/978-3-319-51662-2_6

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329–335.

doi: 10.1037/h0062491

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, *48*(6), 601–612. doi: 10.1111/1460-6984.12061

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., & Babadi, B. (2018). Real-Time Tracking of Selective Auditory Attention From M/EEG: A Bayesian Filtering Approach. *Frontiers in Neuroscience*, *12*. doi: 10.3389/fnins.2018.00262

Mirkovic, B., Bleichner, M. G., De Vos, M., & Debener, S. (2016). Target Speaker Detection with Concealed EEG Around the Ear. *Frontiers in Neuroscience*, *10*. doi: 10.3389/fnins.2016.00349

Mirkovic, B., Debener, S., Jaeger, M., & Vos, M. D. (2015). Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications. *Journal of Neural Engineering*, *12*(4), 046007. doi: 10.1088/1741-2560/12/4/046007

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1591), 919–931. doi: 10.1098/rstb.2011.0355

Moore, T. (1981). Voice communication jamming research. In *Advisory Group for Aerospace Research and Development Conference Proceedings.* Citeseer.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60. doi: 10.1080/17470215908416289

Myers, B. R., Lense, M. D., & Gordon, R. L. (2019). Pushing the Envelope: Developments in Neural Entrainment to Speech and the Biological Underpinnings of Prosody Perception. *Brain Sciences*, *9*(3). doi: 10.3390/brainsci9030070

Nees, M. A. (2016). Have we forgotten auditory sensory memory? Retention intervals in studies of nonverbal auditory working memory. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.01892

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE™//A lexical database for contemporary french : LEXIQUE™. *L'Année psychologique*, *101*(3), 447–462. doi: 10.3406/psy.2001.1341

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85–103. doi: 10.1016/j.wocn.2005.10.004

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085–1099. doi: 10.1121/1.408469

Nogueira, W., Dolhopiatenko, H., Schierholz, I., Büchner, A., Mirkovic, B., Bleichner, M. G., & Debener, S. (2019). Decoding Selective Attention in Normal Hearing Listeners and Bilateral Cochlear Implant Users With Concealed Ear EEG. *Frontiers in Neuroscience*, *13*. doi: 10.3389/fnins.2019.00720

Obleser, J., & Kayser, C. (2019). Neural Entrainment and Attentional Selection in the Listening Brain. *Trends in Cognitive Sciences*, *23*(11), 913–926. doi: 10.1016/j.tics.2019.08.004

Olsen, W. O., Van Tasell, D. J., & Speaks, C. E. (1997). Phoneme and word recognition for words in isolation and in sentences. *Ear and Hearing*, *18*(3), 175.

O'Sullivan, J. A., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of neural engineering*, *14*(5), 056001. doi: 10.1088/1741-2552/aa7ab4

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., ... Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. doi: 10.1093/cercor/bht355

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., ... Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*, *10*(1). doi: 10.1371/journal.pbio.1001251

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex (New York, NY)*, *23*(6), 1378–1387. doi: 10.1093/cercor/bhs118

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. doi: 10.1121/1.1906875

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497–1503. doi: 10.1111/j.1460-9568.2012.08060.x

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual Organization of Sound Begins in the Auditory Periphery. *Current Biology*, *18*(15), 1124–1128. doi: 10.1016/j.cub.2008.06.053

Purves, D. (Ed.). (2018). *Neuroscience* (Sixth edition ed.). New York Oxford: Oxford University Press, Sinauer Associates is an imprint of Oxford Universitiy Press.

Puvvada, K. C., & Simon, J. Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. *Journal of Neuroscience*, *37*(38), 9189–9196. doi: 10.1523/JNEUROSCI.0938-17.2017

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria.

Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2018). Neural Entrainment to Speech Modulates Speech Intelligibility. *Current Biology*, *28*(2), 161-169.e5. doi: 10.1016/j.cub.2017.11.033

Rong, X. (2016). Word2vec Parameter Learning Explained. *arXiv:1411.2738 [cs]*.

Sadek-Khalil, D. (1997). *Apport de la linguistique à la pédagogie*. Montreuil, France: Ed. du Papyrus.

Salamé, P., & Baddeley, A. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 150–164. doi: 10.1016/S0022-5371(82)90521-7

Scott, S. K., & McGettigan, C. (2013). The neural processing of masked speech. *Hearing research*, *303*, 58–66. doi: 10.1016/j.heares.2013.05.001

Sharma, S., Tripathy, R., & Saxena, U. (2016). Critical appraisal of speech in noise tests: A systematic review and survey. *International Journal of Research in Medical Sciences*, *5*(1), 13. doi: 10.18203/2320-6012.ijrms20164525

Shavit-Cohen, K., & Zion Golumbic, E. (2019). The dynamics of attention shifts among concurrent speech in a naturalistic multi-speaker virtual environment. *Frontiers in Human Neuroscience*, *13*. doi: 10.3389/fnhum.2019.00386

Shinn-Cunningham, B., Best, V., & Lee, A. K. C. (2017). Auditory Object Formation and Selection. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party* (pp. 7–40). Cham: Springer International Publishing. doi: 10.1007/978-3-319-51662-2_2

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2019). *Afex: Analysis of factorial experiments*.

Skuk Verena G., & Schweinberger Stefan R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research*, *57*(1), 285–296. doi: 10.1044/1092-4388(2013/12-0314)

Somers, B., Verschueren, E., & Francart, T. (2018). Neural tracking of the speech envelope in cochlear implant users. *Journal of Neural Engineering*, *16*(1), 016003. doi: 10.1088/1741-2552/aae6b9

Søndergaard, P. L., & Majdak, P. (2013). The Auditory Modeling Toolbox. In J. Blauert (Ed.), *The Technology of Binaural Listening* (pp. 33–56). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-37762-4_2

Søndergaard, P. L., Torrésani, B., & Balazs, P. (2012). THE LINEAR TIME FREQUENCY ANALYSIS TOOLBOX. *International Journal of Wavelets, Multiresolution and Information Processing*, *10*(04), 1250032. doi: 10.1142/S0219691312500324

Sur, S., & Sinha, V. K. (2009). Event-related potential: An overview. *Industrial Psychiatry Journal*, *18*(1), 70–73. doi: 10.4103/0972-6748.57865

Teoh, E. S., Cappelloni, M. S., & Lalor, E. C. (2019). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience*, *50*(11), 3831–3842. doi: 10.1111/ejn.14510

Teoh, E. S., & Lalor, E. C. (2019). EEG decoding of the target speaker in a cocktail party scenario: Considerations regarding dynamic switching of talker location. *Journal of Neural Engineering*, *16*(3), 036017. doi: 10.1088/1741-2552/ab0cf1

Theunissen, F., David, S., Singh, N., Hsu, A., Vinje, W., & Gallant, J. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, *12*(3), 289–316. doi: 10.1080/net.12.3.289.316

Turner, R. E., Walters, T. C., Monaghan, J. J. M., & Patterson, R. D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *The Journal of the Acoustical Society of America*, *125*(4), 2374–2386. doi: 10.1121/1.3079772

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). itsadug: Interpreting time series and autocorrelated data using GAMMs.

Van Noorden, L. (1975). *Temporal coherence in the perception of tone sequences* (Unpublished doctoral dissertation). Institute for Perceptual Research, Eindhoven, the Netherlands.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *JARO: Journal of the Association for Research in Otolaryngology*, *19*(2), 181–191. doi: 10.1007/s10162-018-0654-z

Verschueren, E., Somers, B., & Francart, T. (2019). Neural envelope tracking as a measure of speech understanding in cochlear implant users. *Hearing Research*, *373*, 23–31. doi: 10.1016/j.heares.2018.12.004

Vestergaard, M. D., Fyson, N. R. C., & Patterson, R. D. (2009). The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *The Journal of the Acoustical Society of America*, *125*(2), 1114–1124. doi: 10.1121/1.3050321

Vestergaard, M. D., Ives, D. T., & Patterson, R. D. (2009). The advantage of spatial and vocal characteristics in the recognition of competing speech. *Proceedings of the International Symposium on Auditory and Audiological Research*, *2*, 535–544.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392–393. doi: 10.1126/science.167.3917.392

Wilsch, A., Neuling, T., Obleser, J., & Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. *NeuroImage*, *172*, 766–774. doi: 10.1016/j.neuroimage.2018.01.038

Wilsch, A., & Obleser, J. (2016). What works in auditory working memory? A neural oscillations perspective. *Brain Research*, *1640*, 193–207. doi: 10.1016/j.brainres.2015.10.054

Wong, D. D. E., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., & de Cheveigné, A. (2018). A comparison of regularization methods in forward and backward models for auditory attention decoding. *Frontiers in Neuroscience*, *12*. doi: 10.3389/fnins.2018.00531

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (Second ed.). Chapman and Hall/CRC. doi: 10.1201/9781315370279

Yu, T.-L. J., & Schlauch, R. S. (2019). Diagnostic precision of open-set versus closed-set word recognition testing. *Journal of Speech, Language, and Hearing Research*, *62*(6), 2035–2047. doi: 10.1044/2019_JSLHR-H-18-0317

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., ... Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party". *Neuron*, *77*(5), 980–991. doi: 10.1016/j.neuron.2012.12.037

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. doi: 10.1016/j.cub.2017.11.071

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R.* New York, NY: Springer New York. doi: 10.1007/978-0-387-87458-6

APPENDICES

List of Figures in Appendices

List of Tables in Appendices

Appendix A

Supplementary analyses

A.1 Experiments 1 and 2: recency effect

To analyse the recency effect, the duration (in seconds) between the last keyword and the end of the story was added to the generalized linear mixed models (gLMM). Equation A.1 indicates the final model for experiment 1 and equation A.2 indicates the final model for experiment 2 :

$$score \sim duration + presentation * voice + (presentation + voice + duration \mid subject) \tag{A.1}$$

$$score \sim duration + voice + (voice * duration \mid subject) \tag{A.2}$$

Results (see Table A.1 and Table A.2) show that participants had better scores when the last target keyword is close to the end of the sentence.

Table A.1 gLMM coefficients for the distance between the two voices (centered on 0), the stimulus presentation, their interaction, and the duration between the third keyword and the end as fixed factors for experiment 1. β and SE are the estimated value of the coefficient and its standard error. The Wald z value and its associate p-value are the statistics of the coefficient.

Fixed effects	Coefficients		Statistics	
	β	*SE*	z	p
Intercept	4.08	0.27	15.04	$< .001$
Duration (recency)	-0.32	0.07	-4.66	$< .001$
Presentation	1.1	0.37	2.98	$< .01$
Voice	4.2	0.52	8.03	$< .001$
Presentation × Voice	-3.6	0.69	5.2	$< .001$

Table A.2 gLMM coefficients for the distance between the two voices (centered on 0) and the duration between the third keyword and the end as fixed factors for experiment 2.

Fixed effects	Coefficients		Statistics	
	β	*SE*	z	p
Intercept	2.15	0.1	21.85	$< .001$
Duration (recency)	-0.27	0.03	-8.11	$< .001$
Voice	2.54	0.16	16.34	$< .001$

A.2 Experiment 3

A.2.1 Comparison to previous behavioral studies

The score estimation for previous and current studies are represented by Figure A.2 for spatialisation and by Figure A.1 for vocal characteristics. From our data, the voice conditions "JND" and "Male" were chosen to match the Female-Female and the Female-Male condition of Ericson and McKinley (2001). Table A.3 shows the voice parameters and the estimation of the score for previous studies (Darwin et al., 2003; Ives et al., 2010; Vestergaard, Fyson, & Patterson, 2009). All the results are with at a TMR of 0 dB. Note that only the third experiment from Darwin et al. (2003) with the shift from a female voice towards a male voice is displayed (in the Table A.3).

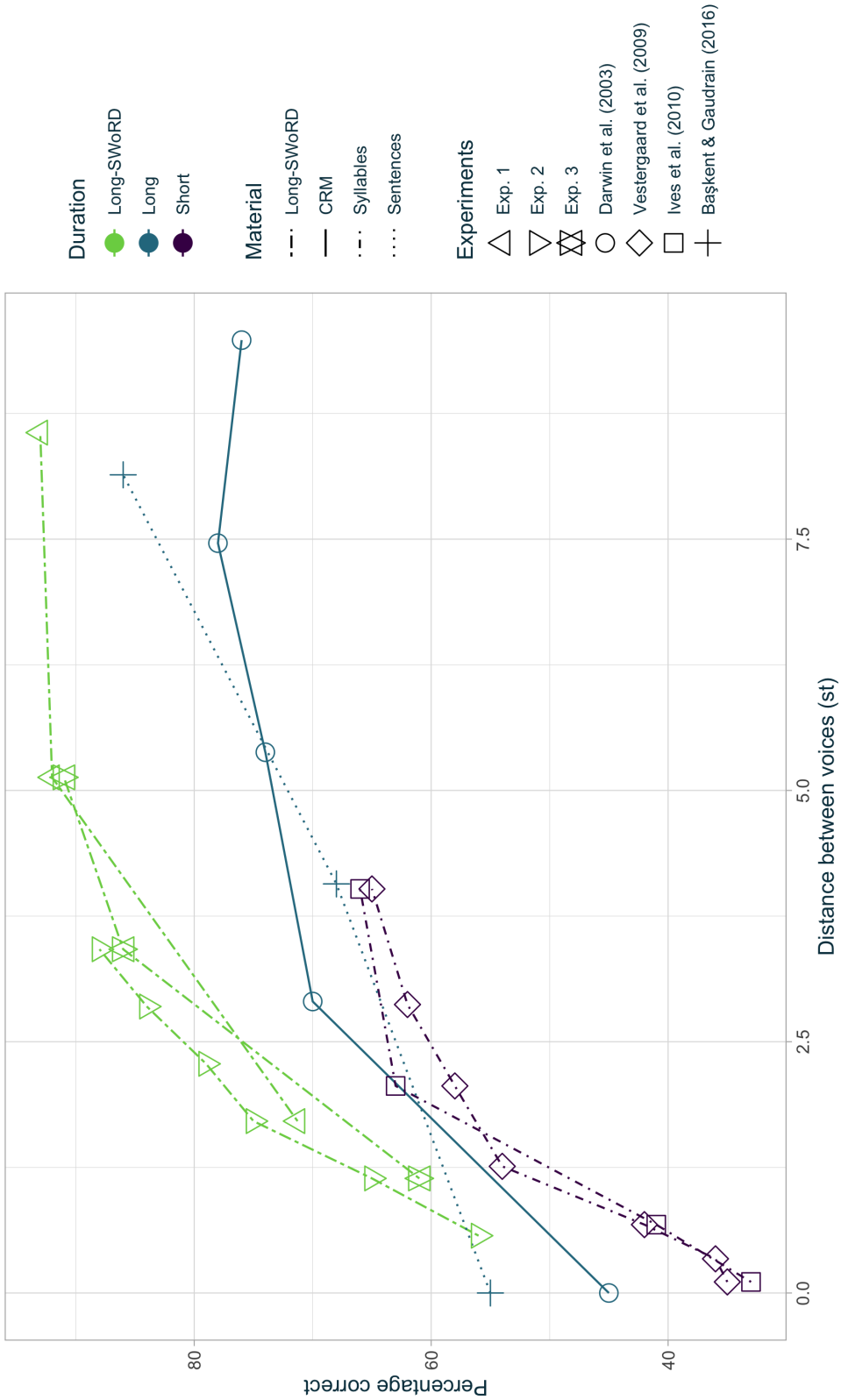

Figure A.1 Scores for previous and current behavioural studies for vocal characteristics.

Table A.3 Voices parameters and estimated scores for previous behavioral studies vocal characteristics

Material	Experience	Voice	ΔF0	ΔVTL	Total distance	Score percentage
Long-SWoRD	Experience 1	JND	-1.6	0.61	1.71	71
		Intermediate	-4.8	1.82	5.13	92
		Male	-8	3.04	8.56	93
	Experience 2	1	-0.53	0.2	0.57	56
		2	-1.06	0.4	1.14	65
		3 (JND)	-1.6	0.61	1.71	75
		4	-2.13	0.81	2.28	79
		5	-2.67	1.01	2.85	84
		6	-3.2	1.21	3.42	88
	Experience 3	Difficult	-1.06	0.4	1.14	61
		Intermediate	-3.2	1.21	3.42	86
		Easy	-4.8	1.82	5.13	91
CRM	Darwin et al. (2003)	Unchanged	0	0	0	45
		1/4 male	-2.81	0.68	2.9	70
		1/2 male	-5.21	1.33	5.38	74
		Almost male	-7.2	1.96	7.46	78
		Male	-9.13	2.57	9.48	76
Sentences	Başkent and Gaudrain (2016)	1	0	0	0	55
		2	4	-0.75	4.07	68
		3	8	-1.5	8.14	86
Syllables	Vestergaard, Fyson, and Patterson (2009)	1	-0.17	0	0.11	35
		2	-0.35	0.17	0.34	36
		3	-0.71	0.17	0.68	42
		4	-1.26	0.51	1.26	54
		5	-2.02	0.68	2.06	58
		6	-2.81	0.84	2.87	62
		7	-3.86	1.12	4.02	65
	Ives et al. (2010)	1	-0.17	0	0.11	33
		2	-0.71	0.17	0.68	41
		3	-2.02	0.68	2.06	64
		4	-3.86	1.12	4.02	66

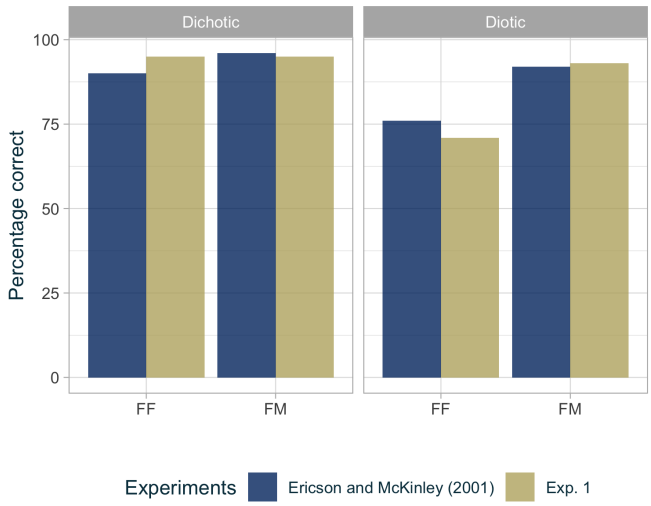

Figure A.2 Scores for previous and current behavioural studies for spatialisation.

A.2.2 Neural tracking

TRF per region of interest

The neural markers of TRFs are reported in the following Table A.4 for the different regions of interest

Stimulus-reconstruction and auditory attention decoding

In a cocktail party tracking, each reconstructed attended stream can be evaluated with the original attended stream ($R_{attended}$) or with original unattended stream ($R_{unattended}$). For instance, if the reconstructed stream is computed with a TRF trained of target envelopes, then the attended stream is the target stream and the unattended is the masker stream. On the other hand, if the reconstructed stream is computed with a TRF trained of masker envelopes, then the attended stream is the masker stream and the unattended is the target stream. Figure A.4 represents the Pearson's correlations for each voice and both $R_{attended}$ and $R_{unattended}$. All the conditions are above chance (multiple paired t-test between the original and the surrogate data with a false discovery rate correction) except the Unattended speech trained with the Target TRF in condition 5.13 $[t(20) = 1.74, p = .96]$.

The comparison between $R_{attended}$ and $R_{unattended}$ leads to the auditory attention decoding. The accuracy of stimulus-reconstruction can also be represented

Table A.4 : TRF Markers latency for each stream and condition (in ms) for each ROI.

ROI	Label	Target			Masker			Difference		
		1.14 st	*3.42 st*	*5.13 st*	*1.14 st*	*3.42 st*	*5.13 st*	*1.14 st*	*3.42 st*	*5.13 st*
Frontal	**"P1"**	0-62	0-47	0-47	0-78	0-125	0-109		0-141	31-94
	P2									172-219
	N4	281-594	312-578	297-578	266-594	297-578	281-578		344-453	
	"P7"									
	"N7"	719-734				672-829		687-750	656-719	
	P8	875-906	796-875				859-906			
PTLeft	**N1**		15-156						0-203	203-266
	P2					203-281				
	N4					344-563		375-531	343-500	
	"P7"			703-750	672-812			687-781		
	"N7"	719-781		765-796						
PTRight	**N1**		109-141			203-281	203- 265		15-296	203-296
	P2									
	N4	296-343	375-516	312-437	375-562	359-594	390-516	422-500	406-468	
	"P7"				703-750	734-796		687-796		
	"N7"	719-765		765-796						

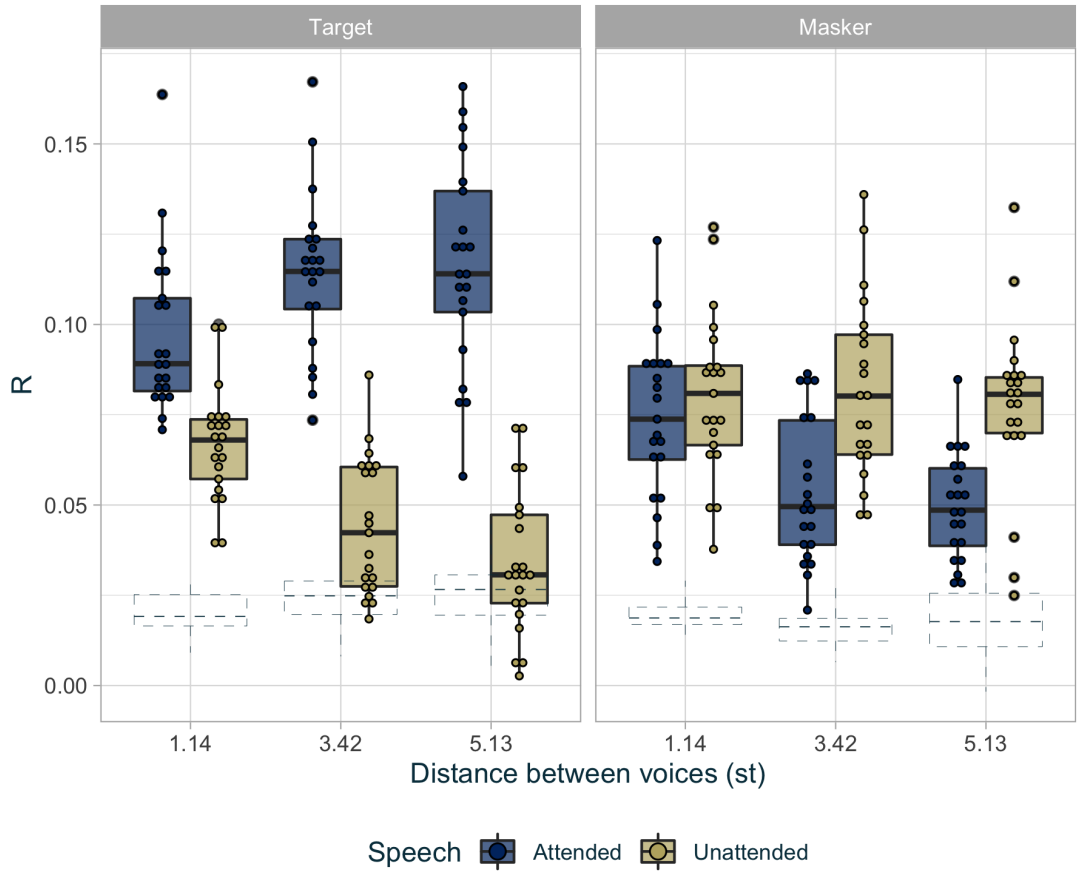

Figure A.3 : Pearson's correlation for each voice when the decoder (TRF) is trained with the target stream (left) or with the masker stream (right). The dots represent the scores for every participant in each condition for the attended stream (in blue) and for the unattended stream (in yellow). The hinges of the boxplot represent the first and the third quartile. Median is represented as a bar in each boxplot. The length of the whiskers is 1.5 interquartile range. The dashed boxplot indicates the chance level.

across time (Figure A.4). The difference between the conditions can be observed around 200 ms.

Build-up

The extra analyses of the build-up effect is shown is Figure A.5

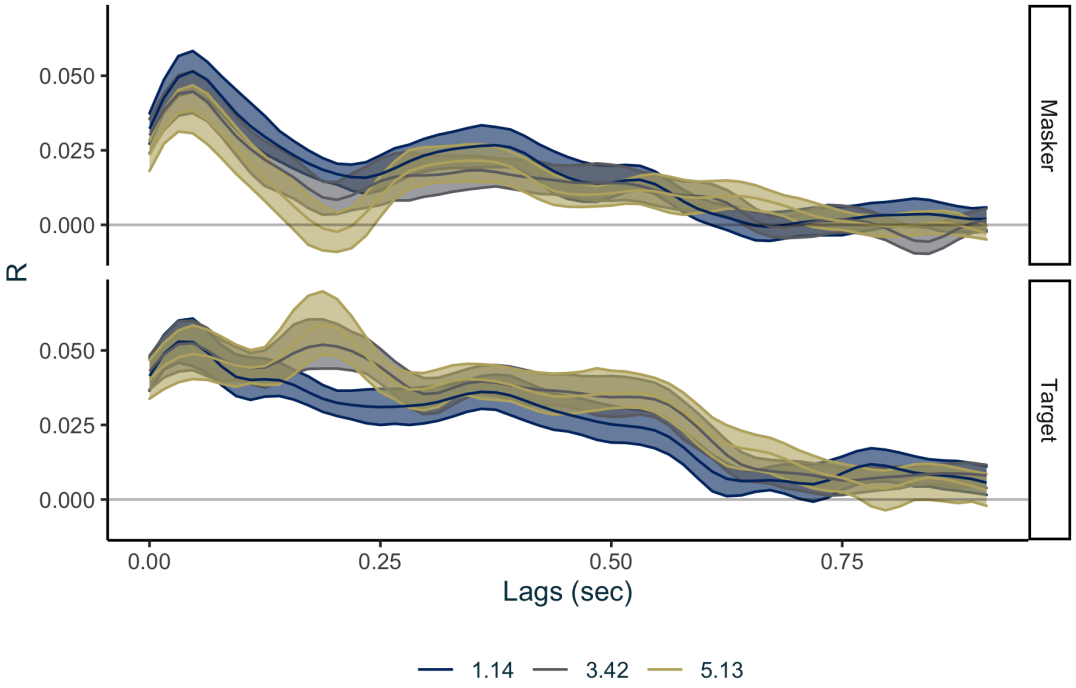

Figure A.4 : Accuracy of stimulus-reconstruction (R) per lag. Confidence intervals (95%) were obtained by bootstrapping the mean across subjects.

A.2.3 Original vs. behavioural stimuli

Stimulus-reconstruction evaluation

The reconstruction for all the behavioural stimuli is represented in Table A.5. When the original target stream was used to train TRFs, the reconstruction quality increased as the task became easier [$R_{1.14} = 0.097; R_{3.42} = 0.114; R_{5.13} = 0.116$] while the masker reconstruction evaluation decreased [$R_{1.14} = 0.074; R_{3.42} = 0.054; R_{5.13} = 0.05$]. When behavioural stimulus-reconstructions were compared with multiple t-tests to the original target voice reconstructions, a pattern emerged very quickly. In general, when extraneous words were filled up with the mixture, the target or even the masker stream, the behavioural stimuli had similar performances than the original target stream. Reconstruction was even significantly better in the most difficult condition (see Table A.5). Conversely, when noise or a different speech stream was used when subjects chose extraneous words, reconstruction performance in the most difficult condition was better but the difference was not significant.

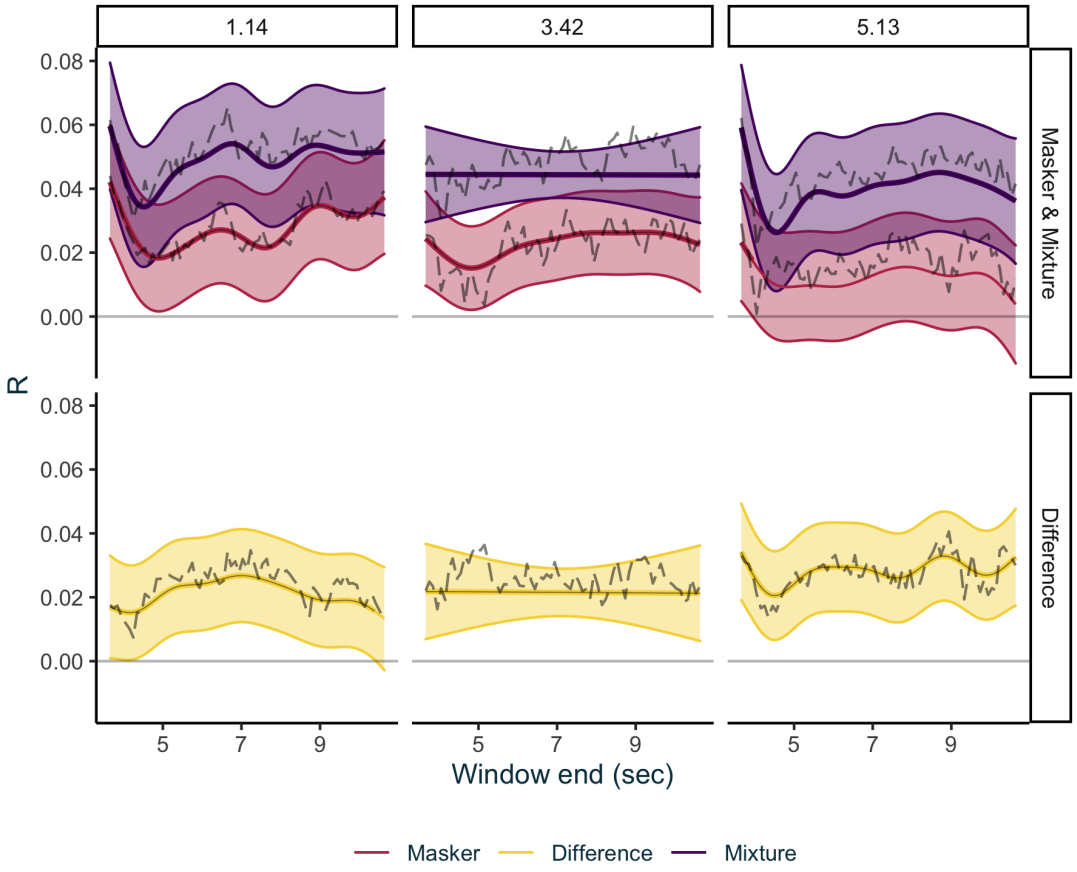

Figure A.5 : Stimulus-reconstruction (R) over time per condition for the masker stream (in rose) and the mixture stream (in purple). The difference between the masker and the mixture streams is shown at the bottom in yellow. Thick lines indicate a response different from zero with 95% confidence intervals. The dashed lines indicate the raw data averaged (before the smoothing of the GAM).

Table A.5 : Behavioural envelopes reconstruction evaluation means. Means that are in bold perform better than the original target while means that are in italic perform worse than the original target

		Sections			Target windows		
Extraneous	**Switch (sec)**	**1.14**	**3.42**	**5.13**	**1.14**	**3.42**	**5.13**
Mixture	1	**0.108**	0.112	0.116	**0.107**	0.113	0.116
	2	**0.108**	0.112	0.116	**0.109**	0.113	0.116
	3	**0.109**	0.113	0.116	**0.109**	0.113	0.116
Target	1	**0.104**	0.11	0.116	**0.104**	0.112	0.116
	2	**0.105**	0.11	0.116	**0.105**	0.112	0.116
	3	**0.105**	0.11	0.116	**0.105**	0.112	0.116
Masker	1	**0.106**	0.109	0.113	**0.106**	0.112	0.114
	2	**0.107**	0.109	0.113	**0.108**	0.111	0.114
	3	**0.107**	0.11	0.113	**0.108**	0.111	0.114
Other speech	1	0.094	*0.103*	*0.111*	0.099	*0.108*	0.114
	2	0.097	*0.103*	0.112	0.01	*0.106*	0.114
	3	0.096	*0.103*	0.113	0.01	*0.107*	*0.112*
Noise	1	0.098	*0.104*	0.113	0.102	*0.11*	0.115
	2	0.099	*0.104*	0.113	0.102	*0.108*	0.114
	3	0.1	*0.105*	0.114	0.102	*0.108*	0.114
Only original target		0.097	0.114	0.116			

Auditory attention decoding

In contrast to stimulus-reconstruction, Table A.6 shows that auditory attention decoding performances achieved with behavioural stimuli were no different from those achieved with the original target voice [$AAD_{1.14} = 60.12\%; AAD_{3.42} = 72.32\%; AAD_{5.13} = 76.59\%$]

Table A.6 : Behavioural envelopes AAD means (in %)

		Sections			Target windows		
Extraneous	**Switch (sec)**	**1.14**	**3.42**	**5.13**	**1.14**	**3.42**	**5.13**
Mixture	1	63.1	69.94	75.6	62.5	69.74	76.29
	2	63	70.44	75	63.59	69.44	75.99
	3	63.89	70.34	74.6	63.1	69.74	75.6
Target	1	63.59	69.64	76.19	62.9	70.24	76.69
	2	62.7	69.94	75.99	62.5	69.94	76.79
	3	63.29	69.84	75.79	62.5	69.25	76.98
Masker	1	64.09	70.14	75.6	62.9	69.74	75.89
	2	63.59	70.73	75.2	63.79	69.64	75
	3	63.89	71.33	74.8	63.59	70.04	75.1
Other speech	1	64.09	69.84	76.29	62.7	70.34	77.18
	2	63.59	70.14	75.4	62.7	70.14	76.19
	3	64.19	69.25	75.5	63	69.05	76.88
Noise	1	65.28	69.44	75.5	63	69.25	76.79
	2	64.58	69.25	75.79	63.49	69.25	76.88
	3	65.18	69.35	75.5	63	69.84	76.49
Only original target		60.12	72.32	76.59			

Appendix B

Long - SWoRD v1

All the stories come from *Le Charme discret de l'intestin* [The Inside Story of Our Body's Most Underrated Organ] (Enders, Enders, & Liber, 2015)

Table B.1 Target and masker stories for list 1 of the Long-SWoRD test v1

Target story	**Masker story**
La raison la plus souvent invoquée pour justifier une prise d'antibiotiques est "un rhume". À ces mots, n'importe quel microbiologiste a les cheveux qui se dressent sur la tête	Les gènes sont des possibilités. Des informations. Les gènes peuvent nous imposer quelque chose comme ils peuvent nous proposer une compétence. Les gènes, surtout, sont des plans.
Même quand le centre de la vue a été détruit suite à un traumatisme au niveau de l'occiput et qu'on est aveugle, on ne "voit" plus l'araignée, mais on la "ressent" toujours. Notre amygdale est donc fortement impliquée dans l'état anxieux.	Sur la base des produits métabolisés par nos bactéries intestinales, on peut déjà, trois semaines à peine après notre naissance, prédire un éventuel risque accru d'allergies, d'asthme ou de dermatite atopique
Toutefois, à long terme, neutraliser l'acidité est plutôt une mauvaise solution : le suc gastrique est aussi là pour brûler les allergènes et les mauvaises bactéries alimentaires, et il contribue à la digestion des protéines.	Pour les enfants qui sont en deçà de la courbe moyenne de poids, l'effet "remplumeur" de l'ablation des amygdales peut présenter un avantage. En prenant du poids, ils vont se retrouver dans la zone dite normale.
Le mécanisme qui se cache derrière l'intolérance au fructose n'est pas le même que pour le gluten ou le lactose. Les personnes qui souffrent d'une intolérance congénitale ont dans leurs cellules peu d'enzymes permettant l'assimilation du fructose	Au supermarché, plantés devant le rayon frais, nous lisons le mot "probiotique" sur un pot de yaourt. Nous ne savons pas vraiment ce que c'est ni comment cela fonctionne – mais beaucoup d'entre nous en ont au moins entendu parler à la télé
En examinant les êtres humains de plus près, on s'aperçoit que chacun d'eux est une petite planète. Le front est une prairie dégagée, les coudes une terre désertique, les yeux des lacs salés et l'intestin une forêt	Les vaisseaux lymphatiques peuvent nous paraître un peu malingres. C'est que leurs parois, à l'inverse de celles des vaisseaux sanguins, ne sont pas pourvues de gros muscles. Le plus souvent, les vaisseaux lymphatiques se contentent d'utiliser la pesanteur
Notre principale protection est la propreté. Nous faisons attention quand il s'agit d'aliments crus, nous n'embrassons pas n'importe quel inconnu, nous lavons à grands seaux d'eau chaude nos agents pathogènes. Mais la propreté n'est pas toujours ce que nous croyons qu'elle est.	Maligne, notre petite poche digestive fait ainsi le tri entre ce qui doit être malaxé et ce qui peut continuer son chemin plus rapidement. Avec son petit air de guingois, notre estomac héberge surtout deux savoir-faire différents.
Dans les supermarchés américains, le sucre entre dans la composition d'environ 80 % des produits proposés. D'un point de vue évolutif, on peut dire que notre corps vient juste de découvrir le placard où maman cache bonbons et friandises	Les plantes fabriquent des antibiotiques qui fonctionnent depuis des siècles sans générer de résistances (rappelons que les champignons comme ceux utilisés pour fabriquer la pénicilline n'entrent pas dans la même catégorie que les plantes, mais dans celle des opisthocontes, comme les êtres humains).
La satiété est généralement signalée par deux entités : d'un côté, le cerveau, et, de l'autre, le reste du corps. À ce stade-là, déjà, ça peut tourner au vinaigre : les gènes de la satiété peuvent par exemple être défaillants chez les personnes en surpoids,	D'autres bactéries se déguisent : les antibiotiques ne reconnaissent plus leurs parois et s'abstiennent alors de les massacrer. D'autres encore utilisent leur faculté de clivage et fabriquent des outils avec lesquels elles vont pouvoir décomposer les antibiotiques
Quand on regarde la Terre depuis l'espace, on ne nous voit pas, nous, les êtres humains. On reconnaît la Terre – un point lumineux parmi d'autres points lumineux sur fond de ténèbres. En se rapprochant, on constate que les êtres humains peuplent des endroits très différents de la planète.	Ce qu'elles aiment, nos bactéries, ce sont les aliments qui arrivent au gros intestin sans avoir été digérés et qu'elles peuvent alors consommer. Mais oh surprise : les pâtes et le pain de mie ne figurent pas sur la liste de leurs aliments préférés
Il y a cent trente ans, l'Europe découvrait que des bactéries étaient à l'origine de la tuberculose. C'était la première fois qu'on prenait vraiment conscience d'elles – comme une menace dangereuse et qui plus est invisible. De nouvelles réglementations furent alors introduites en Europe	les responsables, ce sont les nerfs. Ce sont eux qui régulent les muscles. Si les nerfs optiques n'ont pas repéré la marche, les nerfs des jambes seront mal informés, les jambes vont avancer comme s'il n'y avait pas d'obstacle, et patatras, vous vous retrouvez par terre.
Il ne faut pas beaucoup de temps aux enzymes pour venir à bout des glucides qui entrent dans la composition d'une tranche de pain de mie blanc. Quand il s'agit de pain complet, en revanche, la digestion est plus lente. Le pain complet renferme des glucides particulièrement complexes qui doivent être décomposés progressivement.	Mais au début du XX siècle, les choses changent : les dermatologues allemands exigent "un bain hebdomadaire pour chaque Allemand" ; les grandes entreprises lancent des campagnes sanitaires, font construire des bains pour leurs employés et leur fournissent gratuitement du savon et des serviettes.
nous prenons conscience du fait que la plupart des bactéries sont inoffensives – et mêmes utiles. Certains paramètres ont déjà été décrits scientifiquement. Notre microbiote intestinal peut peser jusqu'à deux kilos et héberge environ 100 billions de bactéries.	Cela vaut donc le coup de faire l'expérience et de supprimer temporairement les aliments riches en glutamate. Pour mener à bien cette entreprise, vous aurez besoin d'une paire de lunettes grossissantes qui, au supermarché, vous permettront de déchiffrer les pattes de mouche de la liste des ingrédients

Table B.2 Keywords and duration for list 1 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
justifier	rhume	microbiologiste	informations	imposer	compétence	tofu	viande	base	11
traumatisme	aveugle	amygdale	bactéries	naissance	asthme	Francfort	scanner	chatouillait	13
acidité	brûler	contribue	courbe	ablation	avantage	installée	endroit	mange	13
fructose	lactose	enzymes	rayon	probiotique	entendu	aspirine	maux	bénéfique	14
planète	prairie	lacs	parois	sanguins	muscles	poches	vaches	distances	14
protection	embrassons	seaux	tri	malaxé	estomac	génome	espoirs	savon	14
sucre	évolutif	placard	antibiotiques	champignons	pénicilline	protecteur	bactéries	candidats	15
entités	vinaigre	défaillants	déguisent	massacrer	outils	sodium	milligrammes	cancers	15
espace	ténèbres	endroits	bactéries	consommer	pain	catégorie	billes	parcours	16
tuberculose	menace	réglementations	régulent	marche	obstacle	laver	diluées	vinaigre	16
enzymes	blanc	complexes	dermatologue	sanitaires	savon	maladie	médecine	évidence	17
inoffensives	scientifiquement	billions	glutamate	lunettes	mouche	minces	dodues	souris	18

Table B.3 Target and masker stories for list 2 of the Long-SWoRD test v1

Target story	**Masker story**
Une psychothérapie vraiment efficace est l'équivalent d'une séance de kiné pour nos nerfs. Elle dénoue les tensions et nous enseigne d'autres mouvements sains – au niveau neuronal	il nous faut toujours avoir sous la main – ou plutôt sous la langue – une petite équipe de joyeux microbes. Les plus sympathiques des bactéries bucco-dentaires ne sont pas exterminées par la salive,
L'odorat est l'un de nos sens les plus fondamentaux. À la différence de ce qui se passe pour le goût, l'ouïe ou la vue, les impressions olfactives ne sont pas contrôlées sur le chemin qui les conduit à notre conscient.	Comme pour l'intolérance au lactose, ce trouble fonctionnel intestinal existe aussi sous une forme sévère congénitale : la fructosémie (ou intolérance héréditaire au fructose),
plus une plante se sent en danger, plus elle injecte de substances nocives dans ses graines. Et si le blé est ainsi rongé par l'angoisse, c'est parce que ses graines ne disposent que de peu de temps pour pousser et se reproduire.	Il y a dans notre ventre – et non dans notre cerveau – une assemblée de bactéries qui, quand elle a été mise au régime les trois jours précédents, votera à l'unanimité un réapprovisionnement en hamburger.
Les salmonelles se multiplient beaucoup plus vite que les cheveux, c'est un premier point. Dès que la température dépasse les 10 °C, la salmonelle sort de son hibernation, s'étire et commence à se développer.	Arrivées à ce stade d'insignifiance, les miettes abandonnent la partie de badminton contre les parois de l'estomac et préfèrent aller faire des glissades : au bout de l'estomac, elles disparaissent par un petit trou comme si c'était la bonde de la baignoire.
Le muscle le plus puissant du corps humain se trouve être celui de la mâchoire, tandis que celui de la langue est le muscle strié le plus agile. Quand ils travaillent ensemble, ces deux-là font des miracles en matière de puissance de broyage et d'agilité	La secrétaire au brushing impeccable gère sur Internet un élevage illégal de furets. Le guitariste du groupe de métal est aussi le meilleur client de la mercerie, parce que le tricot, c'est zen et ça assouplit les doigts.
L'alimentation occidentale se compose à 90 % de ce que nous mangeons et, pour les 10 % restants, de ce que nos bactéries nous donnent chaque jour à manger. Autrement dit : "Mangez neuf repas, le dixième vous est offert !"	endant des siècles, ils ont adapté leur comportement à notre petite personne. Un être humain sur deux héberge des vers au moins une fois dans sa vie. Certains ne le remarquent pas, chez d'autres, c'est un fléau insupportable dont on n'ose à peine parler
Quand on souffre d'une irritabilité de l'intestin, il y a un dysfonctionnement de la communication entre l'intestin et le cerveau – et cela peut être pesant psychologiquement. Si pesant que les effets sont même visibles sur un scanner cérébral.	Assis confortablement devant la télé à regarder des surfeurs musclés prendre la vague, nous voilà surpris par un éternuement. Pas un instant nous ne pensons alors aux figures spectaculaires que réalisent à cet instant d'autres surfeurs, dans nos narines.
E. coli et sa dangereuse jumelle EHEC font par exemple partie de la même famille. Les différences sont minimes, mais néanmoins perceptibles : E. coli est un sous-locataire banal dans nos intestins, EHEC provoque des hémorragies	Règle n 2 : tout ce qui a été en contact avec de la viande crue ou la coquille des œufs doit être soigneusement lavé à l'eau chaude : la planche à découper, les mains du cuisinier, les couverts, les éponges et, le cas échéant, l'essoreuse à salade.
Certains peuples vivent dans des régions hérissées de grandes villes, d'autres sont disséminés à travers de grands espaces presque vierges. Certains vivent dans les paysages glacés du Nord, d'autres dans la forêt vierge ou aux portes du désert	Pour assurer le dynamisme du transit, les médecins recommandent donc un régime riche en fibres : les fibres alimentaires n'étant pas digérées par les enzymes, elles exercent une pression sur les parois de l'intestin, qui répondent alors à l'identique.
Un bon tiers des enfants nés dans les pays industrialisés occidentaux vient au monde par césarienne, en toute élégance : pas de chairs meurtries dans la filière pelvienne, pas d'effets secondaires peu ragoûtants comme la déchirure du périnée ou l'expulsion du placenta	Une fois la bouchée de gâteau suffisamment mâchée (ce qui lui permet d'accéder au rang de "bol alimentaire"), on passe à la déglutition. La langue attrape une petite portion du bol alimentaire (environ 20 millilitres) et la propulse vers le palais mou
La famille intestinale la plus connue, Bacteroides, est aussi celle qui constitue le gros de la masse. Pros de l'assimilation des glucides, les bactéries Bacteroides ont en outre à leur disposition toute une batterie de plans de construction génétique	Si l'on coupe les voies de communication entre ce système et le cerveau, le spectacle donné dans nos viscères ne s'interrompt pas pour autant et les interprètes continuent de s'activer gaiement pour accomplir le travail de digestion.
Nous racontons parfois à nos enfants des mensonges plus gros qu'eux. Je pense par exemple au mensonge du bonhomme barbu qui, une fois par an, pointe son nez et sa hotte pour offrir des cadeaux aux enfants avant de repartir sur un véhicule à la croisée du tapis volant et de la charrette à bœufs	Ce service attentionné, c'est notre système immunitaire qui nous le rend, avec l'aide de nombreuses petites cellules. Il a sous sa direction des experts spécialisés dans la reconnaissance de l'envahisseur, des tueurs à gages, des chapeliers et des pacificateurs.

Table B.4 Keywords and duration for list 2 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
séance	dénoue	sains	main	équipe	sympathiques	olive	artériosclérose	Alzheimer	11
fondamentaux	vue	chemin	trouble	sévère	héréditaire	expérience	participants	comportement	12
nocives	angoisse	pousser	ventre	régime	réapprovisionnement	salée	diététique	supprimant	13
cheveux	températures	hibernation	miettes	glissades	trou	atterrira	mamelon	oxygène	13
mâchoire	agile	miracles	Internet	guitariste	tricot	protéines	spectacle	bulles	14
occidentale	bactéries	repas	siècles	vers	fléau	prix	truie	odorat	14
communication	psychologiquement	scanner	télé	éternuement	figures	énergie	enveloppe	plastique	15
jumelle	différences	locataire	viande	chaude	éponges	japonaise	marines	algues	16
villes	espaces	forêts	transit	alimentaire	pression	grosso modo	perception	mémoire	16
pays	césarienne	périnée	gâteau	déglutition	propulse	matériaux	record	pression	16
connue	assimilation	construction	communication	spectacle	gaiement	dermatologue	sanitaires	savon	17
barbu	hotte	charrette	immunitaire	experts	chapeliers	glutamate	lunettes	mouche	18

Table B.5 Target and masker stories for list 3 of the Long-SWoRD test v1

Target story	Masker story
Quant aux prébiotiques (du grec "pre bios", avant la vie), ce sont des aliments qui atteignent le gros intestin et vont alors nourrir les bonnes bactéries pour qu'elles se développent mieux que les mauvaises	C'est difficile à croire, mais le goût du tofu, qui va d'un arôme neutre à une saveur de noisette, et celui de la viande, prononcé et salé, ont la même base : de nombreux petits acides
On obtient parfois de bons résultats en écoutant de la musique, en s'allongeant sur le côté ou en faisant quelques exercices de relaxation. Pourquoi ? Sans doute parce que ces techniques ont un effet apaisant sur nous.	Avant d'être autorisée à passer dans le sang, une cellule immunitaire doit ainsi participer à un camp d'entraînement, le plus dur qui soit pour les cellules. Elle doit par exemple courir une sorte de "trail"
La cuisson nous permet donc de faire des économies d'énergie, puisque notre estomac n'a plus besoin de déployer l'énergie nécessaire à la dénaturation. En cuisinant, nous délocalisons tout simplement une partie de notre activité digestive.	Le n 14, par exemple, est assis dans un petit wagon, otout en haut des montagnes russes, et lève les bras ; le n 32 fait honneur à une délicieuse salade de opommes de terre à la mayonnaise
Et comme quelques coups de dent ne suffisent pas à déchiqueter Helicobacter et qu'elle est par ailleurs capable de s'adapter, le gros minou et ses descendants ont hérité de notre bactérie. Il y a quand même une justice en ce bas monde	Notre système immunitaire fait la connaissance d'une bande de salmonelles et se dit : "Tiens, je pourrais regarder si je n'ai pas dans mes trésors un ou deux sombreros qui leur iraient." Aussitôt dit, aussitôt fait
Un bébé africain, par exemple, dispose de bactéries qui fabriquent toutes sortes d'outils pour cliver une nourriture très riche en fibres et en végétaux. Chez l'enfant européen, les microbes renoncent en général à cette tâche difficile	Cette zone qu'on appelle aussi l'anneau de Waldeyer fait le tour de la gorge : en bas, nous retrouvons notre champ de bosses (les tonsilles linguales), à droite et à gauche, nous avons les amygdales, dites "tonsilles palatines",
Au pôle Sud, l'environnement est si froid et si stérile que les nouveau-nés n'avaient pas reçu suffisamment de bactéries en héritage. Les températures normales et les germes rencontrés sur le chemin du retour suffirent à les tuer.	Car l'organe le mieux isolé et le plus protégé de tous est loin de tout : il siège dans une enveloppe crânienne osseuse, se love dans d'épaisses méninges et filtre chaque goutte de sang qui veut irriguer ses différentes régions.
Chez les patients atteints du syndrome de l'intestin irritable, le gonflement du ballon déclenche au niveau du cerveau une activité clairement identifiable dans une zone émotionnelle normalement chargée de traiter les sentiments désagréables.	c'est sur cette base que nous élaborons notre ADN, c'est-à-dire notre patrimoine génétique, transmis aux nouvelles cellules fabriquées chaque jour. Et tous les autres êtres vivants, les animaux comme les végétaux, ont recours à ce procédé.
Quand votre chemin croise celui d'une fourmi, rien ne s'oppose donc, du point de vue de cette classification, à ce que vous la saluiez d'égal à égal. Dans l'intestin, les eucaryotes les plus courants sont les levures, qui sont aussi des opisthocontes.	En présence de bactéries responsables de la scarlatine, par exemple, mieux vaut ne pas trop tarder à prendre des antibiotiques. Si la maladie n'est pas combattue à temps, le système immunitaire déboussolé pourrait attaquer sans le vouloir des articulations ou d'autres organes
Souvent, une pression au niveau de l'appendice s'avère douloureuse, alors qu'une pression sur la gauche, étrangement, paraît soulager la douleur. Et dès qu'on retire le doigt à gauche – aïe ! Pourquoi ? Parce que les organes de l'abdomen baignent dans un liquide protecteur.	notre conscient est parfois vexé, voire choqué de se voir traité de la sorte : enfin quoi, il voulait juste siroter tranquillement des téquilas, et voilà le résultat ? Sauf qu'en général, c'est lui le responsable de notre piteux état
L'hygiène fondée sur la peur vise à tout éliminer, tout exterminer. On ne sait pas très bien ce qu'on veut éradiquer, mais on pense en tout cas à quelque chose de méchant, de nuisible. De fait, quand nous faisons le ménage de cette façon, nous éliminons tout	Une autre étude plus récente portant sur des retraités d'Irlande a mis en évidence une nette bipartition : certains des intestins se remettaient très bien de l'antibiothérapie, d'autres en conservaient des séquelles durables. Les raisons de ce phénomène sont encore inconnues
nous faisons face aujourd'hui à une offre de fruits qui, sans globalisation ni transports aériens, n'existerait nulle part sur la planète. En hiver, les ananas des zones tropicales voisinent sur nos étals avec les fraises fraîches des serres hollandaises et quelques figues séchées du Maroc.	John Cryan – un scientifique irlandais – et son équipe sont allés plus loin. En 2011, ils ont nourri la moitié de leurs souris avec une bactérie connue pour ses effets bénéfiques sur l'intestin : le lactobacille L. rhamnosus JB-1
Si nous prenions le temps de dire "Salut !" à chacune de nos bactéries intestinales, nous en serions quittes pour environ trois millions d'années. Notre système immunitaire, lui, ne se contente pas de leur dire "Salut !", il ajoute encore : "Je te trouve très sympa" ou "Je te préfère morte".	L'ascidie commune n'a plus besoin de cerveau dès lors qu'elle s'est établie en un lieu fixe. L'époque du mouvement est révolue pour elle, et le cerveau n'a plus de raison d'être. Penser sans mouvement est moins efficace que d'avoir un siphon oral capable d'aspirer du plancton

Table B.6 Keywords and duration for list 3 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
grec	nourrir	bactérie	tofu	viande	base	main	équipe	sympathiques	11
musique	relaxation	techniques	sang	entraînement	courir	courbe	ablation	avantage	13
économies	déployer	cuisinant	wagon	bras	salade	contient	odeur	grec	13
déchiqueter	minou	justice	connaissance	trésors	sombreros	cerveau	information	marine	13
africain	fibres	européen	anneau	bosses	amygdales	fraise	prodigieusement	grain	14
froid	bactéries	germes	isolé	osseuse	sang	Internet	guitariste	tricot	14
syndrome	émotionnelle	sentiments	ADN	cellules	animaux	premier	envergure	défendre	15
fourmi	intestin	levure	scarlatine	antibiotiques	articulations	classait	terme	visualiser	16
appendice	étrangement	aïe	choqué	téquilas	responsable	viande	chaude	éponges	16
exterminer	éradiquer	ménage	Irlande	intestins	phénomène	antithèse	végétariens	viande	16
globalisation	ananas	fraises	irlandais	souris	intestin	maladie	palmier	fabriquer	17
temps	millions	sympa	fixe	penser	siphon	assimilés	vaisseaux	surface	18

Table B.7 Target and masker stories for list 4 of the Long-SWoRD test v1

Target story	Masker story
les protéines présentes dans les végétaux ne sont pas les mêmes que dans les produits animaux et, souvent, les plantes utilisent une quantité d'acides aminés si réduite qu'on qualifie leurs protéines d'incomplètes.	Matin, midi et soir, nous ne leur offrons plus que des sandwichs sous vide, de la nourriture compartimentée dans de petits plateaux en plastique ou de drôles d'épices inconnues.
Ces guerriers ne se nourrissaient quasiment que de viande et buvaient du lait comme on boit de l'eau, mais, étrangement, cette surconsommation de graisses animales n'entraînait pas de hausse des taux de lipides dans le sang.	Même principe pour le déodorant : en asséchant les aisselles, il les rend moins accueillantes pour les bactéries – et les odeurs sont moins fortes. Le séchage, c'est quand même une belle invention.
Les souris qui présentent des traits dépressifs ne nagent pas bien longtemps. Elles s'immobilisent régulièrement. Dans leur cerveau, les signaux inhibiteurs sont apparemment beaucoup mieux transmis que les impulsions motivantes et stimulatrices	La vitamine spécifique de cet entérotype – qui contient elle aussi du soufre et dégage une odeur prononcée –, c'est la thiamine (du grec theion, "soufre"), également appelée vitamine B1.
Un chat ne peut avoir des toxoplasmes qu'une fois dans sa vie – c'est le moment où il est dangereux pour nous. Les vieux matous, pour la plupart, ont déjà passé le cap de la toxoplasmose et ne représentent alors plus de danger	Cette partie du palais fonctionne comme un interrupteur : il suffit d'appuyer dessus pour mettre en marche le programme de déglutition. La bouche est alors verrouillée (à cette étape, déjà, courant d'air et digestion ne font pas bon ménage).
Le biologiste aime les choses bien ordonnées. Et l'on peut ordonner la Terre comme on met de l'ordre dans son bureau. Pour commencer, on fourre tout dans deux grands tiroirs : le vivant dans un tiroir, le non-vivant dans l'autre.	Mais si la quantité de fructose que nous avons dans le ventre est trop importante pour être assimilée dans son ensemble, nous nous en débarrassons et perdons du même coup celui qui s'est accroché à ses baskets : le tryptophane.
Nombre des aliments autrefois conservés grâce aux bactéries sont aujourd'hui conservés avec du vinaigre – comme la plupart des cornichons. Parfois, on fait fermenter l'aliment avec des bactéries, puis on le fait chauffer pour éliminer les germes	C'est une bonne nouvelle pour tous ceux qui dorment la bouche ouverte, car si nous produisions la nuit les quantités journalières habituelles, soit 1 à 1,5 litre de salive, le résultat sur l'oreiller ne serait pas beau à voir le matin.
le salage a par exemple été une méthode révolutionnaire pour empêcher qu'on ne s'empoisonne avec de la viande avariée. Pendant des siècles, il a donc été d'usage de saler abondamment les viandes et charcuteries pour les conserver	Prenons un sujet relié à tout un tas d'électrodes et posons-lui des questions sur la foi, la personnalité et la moralité ou encore plaçons-le face à une tâche cognitive exigeante, et les scanners révéleront qu'on s'affaire beaucoup dans cette région du cerveau.
Un micro-organisme est particulièrement bien adapté à notre intestin quand il aime l'architecture de nos cellules intestinales, supporte bien le climat et apprécie la cuisine qu'on y sert. Ces trois facteurs diffèrent d'un individu à l'autre	Cette conception de la propreté ne peut pas être judicieuse, car plus les standards d'hygiène sont élevés dans un pays, plus il y a d'allergies et de maladies auto-immunes. Il y a trente ans, une personne sur dix environ était allergique à quelque chose
Le tube digestif est l'architecte intérieur de nos entrailles. Il dessine à droite et à gauche des bourgeons qui gonflent de plus en plus jusqu'à devenir nos poumons. Un peu plus bas, il se retourne comme une poche de pantalon pour former notre foie.	La famille Prevotella est un peu l'antithèse de la famille Bacteroides. D'après les études menées, c'est chez les végétariens qu'elle est la plus fréquente, mais on la rencontre aussi chez les personnes qui ont une consommation raisonnable de viande, ou même chez les vrais fans de bidoche
La science a avec le microbiome un problème que connaît bien la génération Google. Quand nous posons une question, six millions de sources nous répondent en même temps. Nous ne pouvons pas nous contenter alors de dire : "Chacun à votre tour, s'il vous plaît !"	Différents tests permettent de savoir si l'on fait partie de cette catégorie : l'un d'entre eux consiste à avaler de petites billes et à se faire photographier aux rayons X par un médecin, qui suivra ainsi à la trace leur parcours dans l'intestin.
l'insula reçoit des informations affectives en provenance de tout le corps. Chaque information est comme un pixel. Avec tous ces pixels, l'insula crée une image. Cette image est importante, car elle nous fournit une carte géographique des sentiments.	La répartition en entérotypes permettra peut-être à l'avenir de déduire de l'appartenance à tel ou tel type d'intestin un certain nombre de propriétés, comme l'assimilation du soja, la solidité des nerfs ou le risque d'être touché par une maladie ou une autre.
Les animaux incapables de vomir doivent recourir à d'autres stratégies pour s'alimenter en toute sécurité. Les rats et les souris "mordillent" leur nourriture. Ils mâchent de minuscules morceaux en guise de test et n'avalent le reste que si cette première bouchée est bien passée.	on peut avoir un rhume ou être un peu fatigué sans pour autant souffrir d'un déficit en biotine. Et ce qui augmente le taux de cholestérol, c'est plus la grosse portion de lardons sur les pâtes à la carbonara que l'œuf à la coque un peu trop mollet qui nous aura délivré une portion d'avidine

Table B.8 Keywords and duration for list 4 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
végétaux	plantes	réduite	sandwiches	nourriture	drôles	renforce	chouette	euro	12
viande	consommation	lipides	déodorant	odeurs	séchage	pays	infections	France	12
dépressifs	signaux	impulsions	contient	odeur	grec	intelligent	combattre	mauvaise	13
dangereux	matous	cap	interrupteur	déglutition	digestion	soja	osseux	Asiatiques	13
Terre	bureau	vivant	ventre	débarrassons	baskets	perte	différences	infectés	14
vinaigre	fermenter	chauffer	dorment	journalières	résultat	Dracula	génétique	urines	14
révolutionnaire	siècles	charcuteries	électrode	moralité	scanners	pharmaceutiques	chaleur	vapeur	15
adapté	climat	facteurs	judicieuse	hygiène	allergique	disparaître	Harry Potter	gâteau	15
entrailles	bourgeons	poumons	antithèse	végétariens	viande	gâteau	optiques	cérébrales	16
problème	Google	millions	catégorie	billes	parcours	choqué	téquilas	responsable	16
affectives	corps	géographique	appartenance	propriétés	soja	chercheurs	explorateurs	habitants	17
stratégies	souris	bouchée	rhume	cholestérol	mollet	fructose	symptômes	soupe	18

Table B.9 Target and masker stories for list 5 of the Long-SWoRD test v1

Target story	**Masker story**
Quand nous partons en voyage, nous avons la tête pleine : nous pensons à emporter nos clefs, nous éteignons le gaz et n'oublions pas de prendre un livre ou de la musique pour divertir notre cerveau	L'acidification a aussi un autre effet : elle entraîne la coagulation des protéines de lait, et le lait se solidifie. Voilà pourquoi le yaourt n'a pas la même consistance.
La probabilité d'être impliqué dans un accident de la route est plus élevée quand on est colonisé par des toxoplasmes – surtout quand l'infection bat son plein, et moins quand elle est dans sa phase latente.	Trente ans plus tard, un autre scientifique allait se lancer dans une aventure tout aussi passionnante. Pour cela, nul besoin de sillonner les mers : il ouvrit simplement la porte de son petit laboratoire éclairé au néon.
Le cahier des charges de l'élevage bio limite quant à lui plus strictement la quantité d'antibiotiques administrée aux animaux. En cas de dépassement, la viande ne pourra plus être vendue comme provenant de l'agriculture biologique ;	Daniel Wolpert est aussi un neuroscientifique qui juge très significatif le comportement des ascidies. Sa thèse est la suivante : la seule et unique raison d'être d'un cerveau, c'est le mouvement.
Se laver trop souvent n'a aucun sens – qu'il s'agisse des mains ou du reste du corps. En éliminant trop souvent le film graisseux protecteur, nous exposons notre peau sans défense à toutes sortes d'agressions extérieures.	Une molécule de sucre, par exemple, atterrira dans une cellule épidermique du mamelon droit. Elle y sera assimilée et brûlée à l'oxygène, ce qui va générer de l'énergie et maintenir la cellule en vie.
Les premières analyses de notre génome bactérien ne peuvent pas être représentées dans des diagrammes à barres ou des camemberts : les premiers schémas des chercheurs qui étudient le microbiome tiennent plus de l'art contemporain que du tableau à trois colonnes.	on organisa des prélèvements sanguins sur les conducteurs impliqués dans des accidents de la route. L'idée était de savoir s'il y avait plus de porteurs de toxoplasmes parmi les froisseurs de tôle que dans le reste de la population non accidentée.
Quand nous nous crémons les mains, nous utilisons plus ou moins la même méthode : nous enfermons les microbes dans une couche de graisse dont ils ne peuvent s'échapper. En nous lavant les mains, nous éliminons cette couche et, avec elle, les bactéries qu'elle renferme.	Si les problèmes surviennent principalement la nuit, la juste déclivité de l'oreiller peut y remédier : 30°, c'est parfait. Les plus bricoleurs s'armeront d'un rapporteur et d'un tas de coussins avant d'aller se coucher,
En réalité, plutôt que l'estomac, c'est surtout l'intestin grêle qui gargouille. Et si nous gargouillons, ce n'est pas non plus parce que nous avons faim. Le seul moment où nous pouvons faire un peu de ménage, c'est entre deux cycles digestifs	Même quand il fut démontré que le taux de cholestérol des Massaïs baissait de 18 % quand ils buvaient du lait caillé plutôt que du lait normal, on continua de chercher la mystérieuse substance lactique. Il faut croire que trop de zèle nuit
En 2011, sans autre prétention que celle de s'amuser un peu, des chercheurs américains ont étudié la flore du nombril. Dans le nombril de l'un des participants, ils ont trouvé des bactéries qu'on ne connaissait jusqu'ici que du littoral japonais.	on peut affirmer que les porteurs de ce type de bactéries Helicobacter ont effectivement une plus grande probabilité d'être touchés par un cancer de l'estomac – mais aussi nettement moins de risques de mourir d'un cancer des poumons ou d'un accident vasculaire cérébral.
Pendant un repas, quand nous tombons sur quelque chose de très dur, c'est comme si nous ordonnions à toute une équipe de footballeurs professionnels de piétiner l'aliment incriminé pour que nous puissions l'avaler. Pour une bouchée de gâteau, cependant, inutile d'exercer la force maximale	Les neuroscientifiques vont se récrier, mais tant pis – grosso modo, on peut résumer le rôle de ces régions comme suit : perception du "moi", gestion des sentiments, moralité, peur, mémoire et motivation
Un peu d'étymologie n'a jamais fait de mal à personne : le terme cholestérol est formé des termes grecs kholê, la bile, et sterros, ferme. La première fois que le cholestérol a pu être mis en évidence, c'était dans des calculs biliaires.	Sans les antidouleurs de notre salive, ce serait pire. En mâchant, nous sécrétons une dose supplémentaire de ces substances salivaires. Du coup, nos maux de gorge semblent s'atténuer après les repas et les petites plaies de la cavité buccale nous font moins mal.
Comme les végétaux, les bactéries peuvent être classées selon leur lieu d'habitation, leur nourriture et leur degré de toxicité. Pour être exact d'un point de vue scientifique, on devrait parler de microbiote (du grec : "petit" et "vie") pour désigner la population de microbes qui nous habitent,	En dépit des récentes avancées de la recherche, il y a encore des médecins pour considérer les patients atteints du syndrome de l'intestin irritable comme des hypocondriaques ou des simulateurs. À l'examen, aucun dommage visible ne peut en effet être repéré au niveau de l'intestin.
La paroi droite de l'estomac est bien plus courte que la gauche, si bien que notre panse se contorsionne pour former une petite poche bancale aux airs de croissant mal cuit. Quant à l'intestin grêle, il promène ses sept mètres de longueur sans trop savoir où cela le mènera	Des chercheurs ont décrit un processus similaire pour le diabète quand il se déclare chez l'enfant ou l'adolescent. Le système immunitaire détruit alors les cellules du corps qui produisent de l'insuline. L'une des causes possibles pourrait être un dysfonctionnement de la communication avec nos bactéries intestinales

Table B.10 Keywords and duration for list 5 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
voyage	gaz	musique	effet	coagulation	yaourt	plats	envie	quantité	11
accident	colonisé	infection	scientifique	mers	laboratoire	énergie	transmettre	poids	12
élevage	animaux	agriculture	scientifique	thèse	cerveau	influer	arbre	options	13
mains	protecteur	agressions	atterrira	mamelon	oxygène	sang	entraînement	courir	13
génome	diagrammes	chercheurs	conducteurs	route	population	arrosage	frère	inspectent	14
graisse	lavant	bactéries	nuit	bricoleurs	coussins	avion	transpirer	sèche	14
estomac	faim	ménage	cholestérol	caillé	lactique	relaxation	cerveau	fermeture	15
chercheurs	participants	littoral	porteurs	probabilité	vasculaire	phénomène	grain	Descartes	15
footballeurs	aliment	gâteau	grosso modo	perception	mémoire	esprits	discorde	existence	16
étymologie	grecs	évidence	mâchant	gorge	plaies	estomac	gastriques	psychosomatique	16
habitation	grec	microbes	recherche	hypocondriaques	dommage	difficultés	poitrine	myocarde	17
panse	croissant	mètres	diabète	insuline	communication	bisous	bécoter	allaitement	18

Table B.11 Target and masker stories for list 6 of the Long-SWoRD test v1

Target story	Masker story
Les cordes vocales sont priées de se taire et l'épiglotte, tel un chef d'orchestre, se dresse majestueusement (le mouvement est perceptible au niveau du cou) tandis que toute la base de la bouche s'abaisse.	ne entreprise a mis au point un cocktail bactérien qu'on peut utiliser comme un produit nettoyant. En se développant, les bactéries du cocktail inodore font reculer celles qui cocotent du bec.
La procédure est si bien automatisée que nous pouvons même déglutir tout en faisant le poirier. Sans se soucier un seul instant de la pesanteur, notre morceau de gâteau descend donc avec grâce le long de notre buste.	Stephen Collins et son équipe ont par exemple mené une expérience très audacieuse : les participants étaient des souris de deux lignées différentes, dont le comportement a été étudié en détail.
Chacun de nous est pourvu d'un gène dédié à la digestion du lactose. Dans de rares cas, il arrive que les problèmes commencent dès la naissance : les nourrissons touchés ne peuvent pas boire de lait maternel sans souffrir aussitôt de violentes diarrhées	Une fois qu'elle s'est installée, elle reste là où elle est quoi qu'il arrive. La première chose que fait l'ascidie quand elle a trouvé l'endroit de ses rêves, c'est qu'elle mange son cerveau. Ben oui, pourquoi pas ?
Quand nos globules affichent en surface les caractéristiques du groupe sanguin A, nous tolérons aussi le sang d'autres personnes de groupe A. C'est pratique : quand on perd du sang suite à un accident de moto ou à une naissance	certains font des gargarismes quotidiens à l'eau salée, d'autres ne jurent que par la choucroute crue vendue dans les magasins de diététique. D'autres encore prétendent qu'on résout définitivement le problème en supprimant les produits laitiers de son alimentation.
Les êtres humains savent en fait depuis longtemps ce que la recherche découvre peu à peu : ce que nous sommes, c'est aussi ce que nous avons dans le ventre. Nous avons les foies ou l'estomac noué quand nous avons peur	Je veux parler du comte Dracula. En Roumanie, son pays d'origine, on recense une mutation génétique dont les symptômes sont entre autres : une allergie à l'ail et au soleil et des urines couleur sang
À chaque inspiration, nos poumons eux-mêmes ne font rien d'autre que d'ingérer des molécules. "Prendre une bouffée d'air" revient donc à prendre une bouchée d'air, à avaler de la nourriture sous forme gazeuse.	Au début de notre vie, nous aimons la douce sensation de satiété, nous crions au désespoir quand la faim nous taraude et nous pleurnichons sous la torture des ballonnements. Les personnes auxquelles nous faisons confiance sont celles qui nous nourrissent,
La plupart des sauces salade vendues en supermarché ou servies dans les restaurants contiennent du sirop de fructose-glucose. Des études ont pu démontrer que ce sirop inhibait certains messagers chimiques chargés de la satiété	Notre système immunitaire devrait être le premier à s'opposer à cette colonisation de grande envergure. Sur sa "liste de choses à faire", on trouve en effet en assez bonne position : défendre le corps contre les intrus.
Les antihistaminiques modernes, eux, ont été beaucoup améliorés ces dernières années et ne se fixent quasiment pas dans le cerveau. Ils contournent ainsi l'un des effets secondaires liés à l'inhibition de l'histamine dans cet organe : la fatigue.	À l'époque où on ne connaissait pas encore bien les bactéries, on les classait dans le règne végétal – d'où le nom de "flore intestinale". Le terme de "flore" n'est donc pas tout à fait correct, mais il nous permet de bien visualiser ce dont il est question.
Toute personne qui souffre d'états anxieux ou dépressifs devrait garder à l'esprit qu'un ventre mal en point peut aussi être à l'origine d'humeurs noires. Il a d'ailleurs parfois de bonnes raisons, par exemple quand il réagit à trop de stress ou à une allergie alimentaire non détectée.	En 2011, par exemple, cent quatorze Canadiens ont consommé deux fois par jour un yaourt spécialement fabriqué pour eux. La bactérie ajoutée était Lactobacillus reuteri, sous une forme particulièrement gastrorésistante.
Pendant des siècles, nous avons concentré nos efforts sur le "grand monde". Nous l'avons arpenté pour le mesurer, nous avons étudié sa faune et sa flore et philosophé sur la vie qui y est possible. Nous avons construit d'énormes engins et nous avons marché sur la Lune.	La dilution domestique, elle, consiste par exemple à laver les fruits et les légumes. La plupart des bactéries contenues dans la terre sont ainsi suffisamment diluées pour ne plus nous nuire. En Corée, histoire de les embêter un peu plus, on ajoute aussi quelques gouttes de vinaigre à l'eau.
Pour expliquer l'apparition des allergies, une autre hypothèse a été avancée : comme la perméabilité de la paroi intestinale peut parfois brièvement augmenter, des résidus de nourriture profitent de l'opportunité et se fraient alors un passage dans les tissus intestinaux et le sang.	Pour travailler avec ces souris, il faut prendre des précautions énormes : un souffle d'air non filtré, et c'est déjà toute une équipe de germes qui s'invite. Grâce à ces souris, les chercheurs peuvent observer ce qui se passe quand un système immunitaire est au chômage technique.
La coloration des urines en rouge indique que la production de sang ne fonctionne pas correctement et que la personne atteinte élimine par cette voie des précurseurs de l'hème. Autrefois, l'explication d'un tel symptôme était simple : s'il pisse rouge, c'est qu'il a bu du sang.	Quel que soit le nombre de bisous qu'on fait au chien du voisin, quand on a souvent la possibilité de bécoter sa maman, on est bien protégé par les microbes maternels. Avec l'allaitement, elle peut aussi favoriserc ertains germes spécifiques de la flore intestinale, comme les bifidobactéries,

Table B.12 Keywords and duration for list 6 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
taire	orchestre	cou	bactérien	nettoyant	inodore	sandwiches	nourriture	drôles	12
automatisée	poirier	gâteau	expérience	participants	comportement	acides	hygiène	propre	12
gène	naissance	maternel	installée	endroit	mange	infection	poulets	terrain	13
globules	caractéristiques	accident	salée	diététique	supprimant	spécificité	protéine	substance	13
recherche	ventre	noué	Dracula	génétique	urines	nuit	bricoleurs	coussins	14
poumons	molécules	nourriture	satiété	torture	confiance	marche	abri	gastrique	15
supermarché	études	chimiques	premier	envergure	défendre	ADN	cellules	animaux	15
améliorés	cerveau	histamine	classait	terme	visualiser	biberon	Australie	adultes	16
anxieux	humeurs	stress	Canadiens	yaourt	bactérie	mâchant	gorge	plaies	16
efforts	flore	marché	laver	diluées	vinaigre	tête	inquiets	ordinateur	16
hypothèse	résidus	passage	précautions	germes	chercheurs	clavier	signature	exemplaires	17
urines	précurseurs	bu	bisous	bécoter	allaitement	enthousiasme	colonisation	symptômes	18

Table B.13 Target and masker stories for list 7 of the Long-SWoRD test v1

Target story	Masker story
Au programme, il peut donc y avoir un peu du pouce gauche d'Élisabeth, la gentille infirmière, un peu du fleuriste qui a tendu à papa le bouquet de roses, ou un peu du chien de papi	Un affaiblissement du système immunitaire après une opération ou un surnombre de germes résistants après une antibiothérapie de longue durée constituent des situations de danger.
Les glucides simples comme la pâte à tarte, le riz ou les pâtes s'engouffrent rapidement vers l'intestin grêle où ils seront digérés, assurant une augmentation rapide de la glycémie	Les muscles lisses ne sont pas soumis au contrôle volontaire. Au microscope, leur aspect diffère de celui des muscles que nous pouvons commander consciemment, et c'est de là, d'ailleurs, qu'ils tiennent leur nom
D'ailleurs, dans de nombreux pays, les êtres humains se nourrissent intuitivement de plats qui complètent les pièces du puzzle : du riz avec des haricots, des pâtes au fromage, du pain avec de l'houmous ou des toasts au beurre de cacahuète	Parmi elles, les rares à accepter de vivre dans les conditions pas vraiment extrêmes qu'offre un laboratoire sont les archées cryophiles – qui aiment le froid. Elles ont un faible pour les congélateurs à moins 80 °C.
Pour donner un coup de propre à l'intérieur de notre corps, nous disposons de trois nettoyants principaux : les antibiotiques nous permettent de repousser les méchants agents pathogènes, tandis que les prébiotiques et les probiotiques favorisent ce qui nous fait du bien	Dans le cadre d'une étude réalisée à Francfort, en Allemagne, des chercheurs ont même passé les cerveaux des participants au scanner pendant qu'une assistante leur chatouillait les parties génitales à l'aide d'une brosse à dents
Chez les enfants nés par césarienne, il faut attendre des mois ou plus avant que la population bactérienne de l'intestin se normalise. Les trois quarts des nouveau-nés qui attrapent des germes typiques du milieu hospitalier sont des enfants nés par césarienne	Avouons-le : les amygdales, celles qui nous valent souvent de passer sur le billard, ne s'y prennent pas de la meilleure façon qui soit. Au lieu de jouer les pistes noires, comme la langue, elles préfèrent former de profonds sillons, les cryptes
Il existe par exemple des produits à base d'huile de moutarde ou de raifort, ou encore d'extraits de camomille et de sauge. Certains sont à même de lutter non seulement contre les bactéries, mais aussi contre les virus.	Pendant un long voyage en avion, nous nous trouvons dans une situation similaire. Pas besoin de transpirer : il suffit d'être exposé assez longtemps à une atmosphère très sèche, qui va discrètement pomper toutes nos réserves d'eau.
Quand on parle d'intolérance au lactose, il ne s'agit en réalité ni d'une allergie, ni d'une véritable incompatibilité. Cependant, comme en cas d'allergie, la nourriture ne peut pas être décomposée complètement en ses différents éléments.	les moins aventuriers pourront se procurer un coussin spécial dans le commerce. Et puis, la position du buste à 30° est excellente pour la circulation sanguine. Notre professeur de physiologie a dû nous le répéter une bonne trentaine de fois
En améliorant la plasticité nerveuse, les antidépresseurs peuvent déconstruire ce type de schémas – un traitement d'autant plus efficace qu'il s'accompagnera d'une psychothérapie de qualité, le risque de revenir à nos mauvaises habitudes étant alors réduit.	Désormais, la proportion maximale de nitrite de sodium pour les charcuteries s'élève à 100 milligrammes (un millième de gramme) par kilogramme de viande. Et depuis, les cas de cancers de l'estomac ont fortement baissé
Des études ont montré que seules des séances de sport vraiment poussées ont un effet mesurable sur les mouvements intestinaux. Si vous n'avez pas l'intention de faire un marathon, vous pouvez donc – pour ce qui est d'améliorer votre transit – laisser tomber la promenade digestive	Champions dans leur catégorie : le tégument (enveloppe de la graine) du plantain des Indes ou, plus goûteux, le pruneau. Tous deux riches en fibres, ils contiennent en outre des substances actives qui véhiculent une plus grande quantité d'eau dans l'intestin,
Les études menées sur des personnes en surpoids ont montré que leur flore intestinale était moins diversifiée et que certains groupes de bactéries notamment spécialisés dans le métabolisme des glucides y étaient majoritaires. Mais pour faire exploser la balance, ça ne suffit pas encore.	Nos cellules immunitaires sont alors soumises à un stress permanent et, à long terme, ce n'est pas bon pour elles. Qu'on ait quatre, sept ou cinquante ans, quand le système immunitaire est hypersensible, il est parfois préférable de dire adieu à ses amygdales.
Toxoplasma gondii est classé parmi les parasites parce qu'il n'a pas choisi de vivre sur un petit lopin de terre quelconque où se nourrir de l'eau et des plantes locales, mais sur un petit lopin d'être humain. Nous l'appelons "parasite" parce que en échange de notre hospitalité, il ne nous offre rien	le cerveau n'est lui aussi qu'un organe. Quand l'insula crée une image du corps, elle englobe donc aussi l'étage du haut. Il y a là quelques divisions intéressantes comme le département de la compassion sociale, celui de la morale ou encore celui de la logique
Pour générer une envie spécifique, il faut pouvoir accéder au cerveau. Et ce n'est pas chose simple. Le cerveau est bien emmitouflé dans d'épaisses membranes, les méninges. Et à l'intérieur, tous les vaisseaux qui le traversent sont enveloppés de tuniques encore plus denses.	À partir de l'âge de sept ans, plus rien ne distingue la flore intestinale des enfants nés par césarienne de celle des enfants nés par voie naturelle. Les phases précoces pendant lesquelles le système immunitaire et le métabolisme peuvent être influencés sont d'ailleurs révolues elles aussi.

Table B.14 Keywords and duration for list 7 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
pouce	fleuriste	chien	opération	résistants	situations	équipe	stationnée	déploie	11
riz	grêle	augmentation	contrôle	microscope	consciemment	génétique	désactivé	humain	12
intuitivement	puzzle	toasts	conditions	laboratoire	congélateurs	remontées	désagréable	marche	13
propre	antibiotiques	probiotiques	Francfort	scanner	chatouillait	bactéries	naissance	asthme	13
population	Nouveau-nés	hospitalier	billard	jouer	profonds	suc	dénaturent	inaperçues	14
huile	camomille	bactéries	avion	transpirer	sèche	impatiente	étrangères	soldats	14
lactose	incompatibilité	décomposée	coussins	buste	physiologie	économies	avion	lézards	15
antidépresseurs	traitement	psychothérapie	sodium	milligrammes	cancers	troncs	impression	réel	15
sport	marathon	transit	pruneau	substances	quantité	stress	cinquante	hypersensible	16
surpoids	diversifié	balance	stress	cinquante	hypersensible	substances	pruneau	quantité	16
classé	plantes	hospitalité	organe	divisions	morale	lésions	chroniques	stimuli	17
accéder	méninges	enveloppés	flore	naturelle	métabolisme	abeilles	carnivores	énergie	18

Table B.15 Target and masker stories for list 8 of the Long-SWoRD test v1

Target story	Masker story
Pour beaucoup, une pomme par jour ne pose aucun problème – s'il n'y avait pas aussi le ketchup sur les frites, le yaourt sucré aux fruits et la soupe en brique qui, tous, contiennent du fructose.	Dans le cadre des expériences sur les souris de laboratoire, certains sujets pesaient 60 % de plus qu'au départ – un résultat auquel les services de restauration du corps ne peuvent pas arriver tout seuls.
Les fibres, c'est bien, mais ça ne sert pas à grand-chose si on ne boit pas suffisamment : sans eau, elles se changent en gros grumeaux fermes. Gorgées d'eau, elles deviennent de petits ballons	Côté blanchisserie, le principe de dilution est en général amplement suffisant. Pour les torchons humides, la plupart des culottes ou le linge de personnes malades, on peut faire monter les enchères jusqu'à 60 °C.
Chaque seconde, nos reins filtrent et nettoient soigneusement notre sang – un peu comme un filtre à café, mais à un tout autre niveau de précision. Sans compter qu'en général, nos reins ne finissent pas à la poubelle aussitôt après usage...	l'effet protecteur du soja, censé faire rempart contre le cancer de la prostate, les maladies vasculaires et les problèmes osseux, est aujourd'hui confirmé. Plus de 50 % des Asiatiques en profitent.
Nettoyer devrait consister à réduire le nombre de bactéries – pas à les éliminer toutes. Même les mauvaises bactéries peuvent être bonnes pour nous tant que notre corps peut les utiliser pour s'entraîner et garder la forme	Avoir l'estomac vide n'empêche pas de vomir, puisque l'intestin grêle peut lui aussi renvoyer son contenu en sens inverse. L'estomac ouvre alors ses portes et laisse passer le contenu de l'intestin grêle vers l'étage supérieur.
L'alcool n'atteint pas en premier les nerfs du cerveau, mais ceux de l'intestin – dans quelle mesure le délassement procuré par un "petit verre de vin au dîner" pourrait-il alors être induit par l'apaisement du cerveau du ventre	Dans notre quotidien aussi, le froid joue un rôle important. Songeons par exemple à la réfrigération des aliments. Le problème, c'est que nos réfrigérateurs sont souvent si pleins que même à basses températures, ils font la joie des bactéries.
Quand nous buvons du vin ou de la vodka, nous savourons en réalité le produit final de la métabolisation des levures, c'est-à-dire l'alcool. Et le travail du petit peuple ne s'arrête pas dans le tonneau de vin, loin de là	Les valeurs ne sont pas élevées au point de devoir mettre en place un traitement comme en cas de plaie importante ou de septicémie, c'est pourquoi on parle d'"inflammation subclinique". Et s'il y a bien quelqu'un qui s'y connaît en matière d'inflammation, ce sont les bactéries
Chez les insectes, le gluten bloque une enzyme digestive importante, si bien qu'une sauterelle qui grignote trop de plants de blé se retrouve avec des crampes d'estomac, Elle laisse donc en paix le reste du champ, et tout le monde est content.	Quant aux techniques de relaxation, elles ont pour effet de réduire le nombre de messages précipités envoyés au cerveau. Si tout se passe bien, la fermeture du sphincter œsophagien inférieur devrait alors être plus stable
nous savons que les tout premiers habitants de notre ventre sont des éléments déterminants pour l'avenir de notre corps tout entier. Sur ce point, les études mettent surtout en évidence l'importance pour le système immunitaire des premières semaines de notre existence,	Nous ne nous faisons pas de bile quand tout va bien. Nous ravalons notre colère, digérons les affronts qui nous sont faits et nos échecs nous laissent un goût amer. Et quand nous sommes émus, nous sommes pris aux tripes.
Chaque amateur de vin décèlera donc un goût légèrement différent – en fonction de ses bactéries. Mais que cela ne nous empêche pas d'écouter attentivement ce que nous raconte notre gentil œnologue. Ce n'est pas tous les jours qu'on trouve quelqu'un pour nous parler de ses microbes avec tant de fierté.	Si, pour un nombre de calories identique, une banane fait moins grossir qu'une barre chocolatée, c'est parce que les glucides d'origine végétale attirent plutôt l'attention des cantinières locales que celle des services de restauration actifs à l'échelle du corps tout entier
Chaque être humain a sa petite collection de bactéries bien à lui, à partir de laquelle on pourrait même établir une empreinte bactérienne de chacun d'entre nous. Si on faisait un prélèvement sur un chien et qu'on analysait les gènes de ses bactéries, on pourrait très certainement retrouver son maître	dans la mesure du possible, on fait bien chauffer la viande et les plats à base d'œufs. Se lever à la fin d'un dîner aux chandelles pour aller mettre sa part de tiramisu au micro-ondes aurait sans doute des conséquences désastreuses sur la qualité de la relation amoureuse,
Notre salive contient une substance antalgique bien plus puissante que la morphine : l'opiorphine. Celle-ci n'a été mise en évidence que récemment, en 2006, par les chercheurs de l'Institut Pasteur. Évidemment, nous ne produisons cette substance qu'en très petites quantités	notre jeune ascidie navigue à travers les mers. Elle cherche son petit coin de paradis. Dès qu'elle a trouvé un rocher qui lui paraît sûr, qui a la température idéale et qui se trouve à proximité de réserves alimentaires, elle pose ses valises.
La salive, c'est du sang filtré, passé au chinois par les glandes salivaires qui retiennent les globules rouges, plus utiles dans nos veines que dans notre bouche. Le calcium, les hormones et les anticorps du système immunitaire contenus dans le sang, en revanche, passent dans la salive.	Les chercheurs ont administré aux rongeurs un mélange de trois antibiotiques différents qui n'agissent que dans l'intestin, éradiquant ainsi la totalité des bactéries intestinales. Ensuite, ils ont transplanté sur les animaux d'une lignée les bactéries intestinales typiques de l'autre lignée,

Table B.16 Keywords and duration for list 8 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
pomme	ketchup	soupe	souris	résultats	restauration	déodorant	odeurs	séchage	12
boit	grumeaux	gorgées	dilution	humide	malades	scientifique	mers	laboratoire	12
soigneusement	café	poubelle	soja	osseux	Asiatiques	bactéries	ressemble	question	13
réduire	bonnes	entraîner	vomir	renvoyer	étage	spécificité	protéine	substance	13
nerfs	verre	apaisement	froid	aliments	températures	rayon	probiotique	entendu	14
vodka	levures	tonneau	traitement	septicémie	clinique	santé	dépressifs	nerveux	14
sauterelle	crampes	champ	relaxation	cerveau	fermeture	cholestérol	caillé	lactique	15
habitants	études	semaines	bile	affronts	émus	cerveau	digestif	barrière	15
vin	œnologue	microbes	banane	végétale	restauration	scarlatine	antibiotiques	articulations	16
collection	empreinte	chien	chauffer	dîner	tiramisu	gâteau	déglutition	propulse	16
antalgique	morphine	Pasteur	mers	rocher	réserves	résidus	bactérie	lait	17
glandes	globules	hormones	rongeurs	totalité	transplanté	chaussure	supermarché	yaourt	18

Table B.17 Target and masker stories for list 9 of the Long-SWoRD test v1

Target story	Masker story
Il n'empêche : notre corps aime tout ce qui est bien sucré, car cela lui évite du travail. Comme les protéines chaudes, ce sucre est assimilé plus rapidement. Et peut être converti tout aussi vite en énergie.	Les adeptes de break dance rapprocheraient ce mouvement de la figure du serpent ou du ver. Les médecins, eux, ont choisi pour décrire ce phénomène le terme a priori moins parlant de péristaltisme,
Les signaux en provenance de l'intestin peuvent arriver dans différentes régions du cerveau, mais pas dans toutes. Ils n'ont par exemple jamais pour destination le cortex visuel, au-dessus de la nuque.	Que nous dégustions un steak, savourions une salade ou soyons trop ivres pour ne pas remarquer que nous mâchouillons une natte en raphia – nos amies Bacteroides savent tout de suite de quelles enzymes elles ont besoin.
Quand la sensibilité au gluten n'est avérée que dans ce type de situations, le patient peut même présenter les signes d'une véritable intolérance. La meilleure chose à faire dans ce cas, c'est de supprimer le gluten pendant un temps.	Sans oublier que nous pouvons aussi influer sur le monde en réfrénant le mouvement. En revanche, quand on est un arbre, par exemple, et qu'on n'a pas le choix entre ces deux options, eh bien, on n'a pas besoin de cerveau.
En 2005, Barry Marshall et John Warren ont reçu le prix Nobel de physiologie ou médecine pour leur découverte de l'implication d'Helicobacter pylori dans les inflammations, les ulcères et le cancer.	On fit des tests avec le lait de vache, mais aussi avec le lait de chamelle et le lait de rate. On parvint à faire baisser le taux de cholestérol – parfois, mais pas toujours : les résultats ne menaient les scientifiques nulle part
les souris stériles sont un peu bizarres. Elles sont souvent hyperactives et, pour des souris, elles se montrent très casse-cou. Par rapport à leurs congénères "habitées", elles mangent plus et mettent plus de temps à digérer	L'être humain, lui, préfère se rendre au supermarché et y acheter de la viande ou du tofu – c'est une façon comme une autre de compenser son incapacité à assimiler les bactéries riches en protéines présentes dans son gros intestin.
Les toxoplasmes se multiplient dans les intestins des chats. Le chat est leur hôte, et tous les autres animaux qui ne servent aux toxoplasmes que de taxis pour se rendre jusqu'au prochain matou sont des hôtes intermédiaires.	dans la première catégorie, on se soucie de sa santé et on surveille très attentivement son alimentation, tandis que dans la deuxième, on commence à en avoir assez de ne plus pouvoir préparer un repas pour des amis sans avoir à passer avant à la pharmacie
Parmi nos collègues vomisseurs, citons les singes, les chiens, les chats, les cochons, les poissons et les oiseaux. En revanche, vous ne verrez jamais vomir une souris, un rat, un cochon d'Inde, un lapin ou un cheval.	les lipides sont capables de concentrer deux fois plus d'énergie par gramme que les glucides ou les protéines. Grâce à eux, nous constituons une enveloppe autour de nos nerfs – un peu comme une gaine en plastique recouvre un câble électrique
Penser intelligemment l'hygiène, c'est autre chose : les recherches en bactériologie marquent l'avènement d'une nouvelle conception de la propreté, dans laquelle il n'y a que peu de place pour l'extermination systématique des petits peuples qui nous entourent	Au cours de son voyage le long de l'intestin grêle, le chyme va disparaître presque en totalité par les parois – un peu comme Harry Potter traversant le mur pour rejoindre le quai 9 3/4. L'intestin grêle saisit notre bout de gâteau à bras-le-corps.
En vieillissant, nous avons tendance à avaler plus souvent de travers : les muscles qui coordonnent le spectacle ne respectent plus aussi bien la chorégraphie, le muscle constricteur supérieur a parfois un temps de retard et l'épiglotte a besoin d'une canne pour se lever.	Les particules de lumière qui rebondissent sur la part de gâteau sont projetées sur la rétine et activent les nerfs optiques. À l'issue d'une petite promenade dans les circonvolutions cérébrales, cette "première impression" est envoyée au cortex visuel
Dans les années 1950, le bain hebdomadaire finit par s'imposer. La classe moyenne prend son bain le samedi – chaque membre de la famille s'immergeant l'un après l'autre dans la même eau. Et dans certains foyers, c'est le père, le travailleur, qui passe en premier	C'est là qu'intervient le réflexe péristaltique. Le premier scientifique à avoir mis en valeur ce mécanisme a isolé un morceau d'intestin, puis y a insufflé de l'air par un petit tuyau – et l'intestin, courtois, a soufflé de l'air en retour
Pour rendre correctement compte des dernières nouvelles de notre planète microbienne, chacun de nous devrait employer au moins une grande agence de presse internationale. Et quand il nous arrive de nous ennuyer, de penser qu'il ne se passe rien dans notre vie, il nous suffit d'y regarder d'un peu plus près	Quand on regarde les chiffres, tout cela ne semble pas si catastrophique : seul 1 % environ des porteurs d'Helicobacter développe un cancer de l'estomac. Mais quand on se rappelle que la moitié de l'humanité est porteuse de ce germe, 1 %, ça fait quand même un sacré paquet de gens.
La globalisation, ce n'est pas seulement quand le bistrot du coin se transforme en McDonald's, c'est aussi ce qui se passe jusque dans nos nombrils. Chaque jour, des milliards de milliards de micro- organismes font le tour de la planète en avion sans payer un seul centime.	En leur injectant un cocktail de bactéries provenant d'autres souris, on peut observer des effets étonnants. Quand on leur administre des bactéries de sujets diabétiques (de type 2), les souris de laboratoire développent rapidement les premiers problèmes de métabolisation des sucres.

Table B.18 Keywords and duration for list 9 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
sucré	travail	converti	mouvement	ver	phénomène	mères	route	litre	12
intestin	régions	visuel	steak	amies	enzymes	calcium	émail	constamment	12
avérée	intolérance	supprimer	influer	arbre	options	vache	rate	scientifiques	13
Nobel	médecine	inflammation	vache	rate	scientifiques	vomir	renvoyer	étage	13
bizarres	casse-cou	mangent	supermarché	tofu	protéine	vapeur	cristaux	emprisonnés	14
multiplient	taxis	matou	santé	deuxième	amis	enceinte	enfant	rapide	14
singes	oiseaux	lapin	énergie	enveloppe	plastique	langue	contre-courant	glandes	15
hygiène	propreté	extermination	disparaître	Harry Potter	gâteau	voitures	rate	chat	15
travers	chorégraphie	épiglotte	gâteau	optiques	cérébrales	parasite	humeur	grignotés	16
hebdomadaire	famille	travailleur	scientifique	mécanisme	courtois	chercheurs	patience	trace	16
planète	presse	ennuyer	catastrophique	cancer	humanité	patientes	emménager	jardins	17
bistrot	nombrils	avion	cocktail	diabétiques	problème	fixe	penser	siphon	18

Table B.19 Target and masker stories for list 10 of the Long-SWoRD test v1

Target story	Masker story
Tandis que nos ancêtres, les chasseurs-cueilleurs, mangeaient chaque année jusqu'à cinq cents variétés de racines, d'herbes et de végétaux, nous nous nourrissons aujourd'hui le plus souvent de dix-sept plantes utiles.	Une fois dans les cellules cérébrales, ces deux acides aminés sont transformés en dopamine et en sérotonine. Vous avez dit "dopamine" ? Mais oui, c'est LE mot magique quand on parle du circuit de la récompense
À en croire les médias, le cholestérol serait mauvais en soi. Rien de plus faux. Trop de cholestérol, ce n'est pas terrible, c'est vrai, mais pas assez de cholestérol, ça ne vaut pas mieux.	Pour que nous puissions humer le délicieux parfum de notre part de gâteau, il faut que certaines molécules surfent sur un courant d'air et soient attirées à l'intérieur de la cavité nasale lors de l'inspiration.
Si la lessive agit contre les taches, c'est parce qu'elle "digère" les substances graisseuses, protéagineuses et glucidiques et les rejette dans les eaux usées pendant que le linge est malaxé dans le tambour. C'est à peu près ce qui se passe dans l'intestin grêle.	Chez l'être humain, par exemple, les yeux envoient au cerveau l'image d'un panneau de signalisation. Chez l'ascidie, les yeux envoient une information sur la densité de circulation sous-marine à cette heure-là.
Notre ventre abrite entre trois et six mètres d'intestin grêle formant des lacets que rien ne maintient. Un tour sur le trampoline ? L'intestin fait des bonds lui aussi. L'avion décolle ? L'intestin est comme nous pressé contre le siège.	Pour renforcer les défenses du maïs face aux nuisibles, on a en effet introduit des gènes qui lui permettent de fabriquer de l'avidine. Quand les nuisibles – ou de naïfs cochons – consomment ce maïs, ils s'intoxiquent.
L'intestin a à sa disposition toute une cohorte de messagers chimiques, de matériaux d'isolation cellulaire et de types de connexion. Il n'y a qu'un autre organe qui offre une telle diversité : le cerveau.	Pour comprendre comment fonctionne ce coude, rappelez-vous le jeu du tuyau d'arrosage. On demande à sa grande sœur ou à son grand frère pourquoi l'eau ne coule plus et, quand ils inspectent l'embout du tuyau, on relâche le coude.
La peur est associée à une partie du cerveau qu'on appelle l'amygdale. Il existe des fibres nerveuses qui relient directement les yeux à cette zone, de sorte qu'en voyant une araignée, nous pouvons immédiatement éprouver de la peur	Dans notre laboratoire, la bibliothèque bactérienne se compose de curieux germes qui ont survécu à des températures de moins 80 °C et qui, une fois décongelés, continuent leur petit train- train habitue
D'après la médecine traditionnelle chinoise, le point P 6 active des méridiens qui vont des bras vers le cœur, détendent le diaphragme et se prolongent jusqu'à l'estomac ou, plus bas, jusque dans le bassin	Les orifices cachés sous la langue, eux, travaillent en continu. Si l'on pouvait entrer par ces petits trous et nager à contre-courant dans le canal excréteur, on arriverait aux glandes salivaires en chef, qui produisent la majeure partie de la salive
Si la puberté est une période troublante pour le cerveau humain, c'est parce que les nerfs sont alors incroyablement plastiques – beaucoup de choses ne sont pas encore établies, tout est possible, rien n'est figé, et les informations fusent dans tous les sens.	le système immunitaire active une enzyme (IDO) pour nous protéger des parasites. Celle-ci résorbe alors en quantité accrue une substance dont les envahisseurs se nourrissent et les force ainsi à entrer dans une phase de somnolence et d'inaction.
Les bactéries ne se contentent pas de décomposer notre repas, elles en profitent aussi pour fabriquer de toutes nouvelles substances. Un chou blanc, par exemple, contient moins de vitamines que la choucroute qu'il deviendra. Les vitamines supplémentaires, ce sont les bactéries qui les fabriquent.	Ces souris sont les animaux les plus propres du monde – naissance stérile par césarienne, cages désinfectées au chlore et alimentation stérilisée à la vapeur. Jamais on ne trouvera dans la nature des animaux ainsi vierges de tout germe
l'Organisation mondiale de la santé recommande de se débarrasser de l'éventuel responsable. Même chose quand on a dans sa famille des cas de cancer de l'estomac, certains lymphomes ou des antécédents de Parkinson : mieux vaut mettre Helicobacter à la porte.	Et le cerveau a aussi une influence sur la sécrétion des sucs. Les nerfs de l'appareil digestif, eux, veillent à ce que l'œsophage ne perde pas le rythme de sa ola et reste bien propre grâce aux milliers de gorgées de salive avalées chaque jour.
Notre intestin veut nous offrir la plus grande surface possible et, pour y parvenir, il se plie en quatre pour nous. Prenons pour commencer les plis visibles à l'œil nu : sans eux, pour que nous puissions bénéficier de la même surface digestive, il nous faudrait un intestin de 18 mètres.	Nous pouvons désormais formuler cette hypothèse : tout comme nous sommes influencés par le grand monde dans lequel nous vivons, le petit monde qui vit en nous nous influence aussi. Et ce qui rend les choses encore plus passionnantes, c'est que ce petit monde n'est pas le même chez chacun d'entre nous.
C'est aussi sur cet effet que misent certains magasins qui pratiquent le "marketing olfactif". Une marque de vêtements américaine utilise même des phéromones sexuelles. À Francfort, en Allemagne, on voit régulièrement des files d'ados faire la queue devant le magasin sombre au parfum envoûtant.	Mais les choses étant ce qu'elles sont, chaque être vivant de taille respectable accueille au moins un autre être vivant qui l'aide et qui, en échange, a le droit de s'installer chez lui. Voilà pourquoi nous avons des cellules dont la surface est très bien adaptée à la fixation des bactéries

Table B.20 Keywords and duration for list 10 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
chasseurs	variétés	végétaux	cérébrales	sérotonine	circuit	contrôle	microscope	consciemment	12
médias	faux	vrai	parfum	molécule	cavité	bactéries	participants	sentiments	12
taches	graisseuses	linge	cerveau	information	marine	miettes	glissades	trou	13
lacet	trampoline	décolle	défenses	fabriquer	cochons	anneau	bosses	amygdales	14
chimiques	connexion	organe	arrosage	frère	inspectent	bibliothèque	températures	décongelé	14
amygdale	yeux	araignée	bibliothèque	températures	décongelé	isolé	osseuse	sang	14
chinoise	méridiens	estomac	langue	contre-courant	glandes	hamac	dessert	viande	15
période	plastiques	figé	enzyme	envahisseurs	somnolence	minutes	génération	microscopique	15
repas	chou	choucroute	propres	stérilisée	nature	courants	polynésiens	argumenter	16
recommande	famille	Parkinson	sécrétion	oesophage	gorgées	chauffer	dîner	tiramisu	16
parvenir	bénéficier	digestive	hypothèse	grand	passionnantes	mers	rocher	réserves	17
marketing	Francfort	parfum	taille	installer	fixation	symptôme	France	systématique	17

Table B.21 Target and masker stories for list 11 of the Long-SWoRD test v1

Target story	Masker story
Règle n 1 : les planches à découper en bois, c'est joli, c'est tendance... mais les bactéries survivent bien mieux dans les fentes et les rainures du bois que sur une planche en plastique.	Quant à savoir pourquoi presque personne ne fait d'allergies aux lardons, la réponse est simple : nous sommes nous-mêmes faits de chair et n'avons généralement aucun problème à la digérer.
Si l'on s'en tient sagement à des œufs bios de poules nourries avec les produits de la ferme, on est moins exposé aux bactéries dangereuses – à moins que l'agriculteur ait un faible pour le poulet surgelé soldé.	Dans le ventre de nos mères, déjà, nous nous entraînons à déglutir et pouvons avaler jusqu'à un demi-litre de liquide amniotique. Si jamais nous faisons une fausse route, inutile d'appeler les secours
Les molécules de sucre ont la capacité de former ensemble des chaînes complexes et perdent alors leur goût sucré : ce sont les glucides que l'on trouve par exemple dans des aliments comme le pain, les pâtes ou le riz	Un bonbon à la menthe nous procure un goût frais, des fenêtres lavées paraissent claires et se glisser dans des draps propres après avoir pris une douche est un avant-goût du paradis. Nous passons volontiers la main sur des surfaces lisses
Faire la connaissance de nos microbes intestinaux n'est pas toujours facile. Ils n'aiment pas vraiment sortir de chez eux. Si on leur installe un petit coin douillet en laboratoire pour pouvoir les observer, ils font grève	Notre corps est un corps intelligent. Entre risque et profit, il sait faire la part des choses : s'il s'agit de combattre un parasite dans le cerveau, eh bien, soit, nous serons de mauvaise humeur
Nous savons aussi que les intestins de Tokyo sont capables de décomposer les algues marines, tandis que ceux de Châtillon-sur-Seine sont moins doués pour ça. Nos bactéries intestinales délivrent des renseignements approximatifs sur notre identité	La situation est tout autre quand la personne infectée est une femme enceinte. Les parasites peuvent s'infiltrer par le sang et parvenir jusqu'à l'enfant. Son système immunitaire ne les connaît pas encore et n'est pas assez rapide pour les neutraliser.
Une stimulation permanente n'est jamais très bénéfique. On s'en aperçoit bien avec les piqûres d'insecte. Quand ça démange en continu, on finit par perdre patience et, pour que les démangeaisons cessent enfin, on se met à gratter jusqu'à ce que ça saigne	nous tétons notre mère, nous nous faisons les dents sur un barreau de chaise, nous embrassons la vitre de la voiture ou le chien du voisin... Tout ce qui parvient dans notre bouche de cette manière pourrait peu après étendre son empire sur nos entrailles
L'une des hypothèses sur l'apparition des allergies a pour point de départ la phase digestive qui se déroule dans l'intestin grêle. Si nous ne parvenons pas à fragmenter une protéine en différents acides aminés, de minuscules morceaux peuvent subsister.	certains rats adoptaient soudain un tout autre comportement. Ils semblaient sans crainte et exploraient tous les recoins de la cage, pénétraient – à l'encontre de leurs instincts – dans la maisonnette marquée par l'urine de chat et y restaient même un certain temps.
Les antibiotiques n'ont pas leur pareil pour tuer les agents pathogènes. Et leurs familles. Et leurs amis. Et les amis de leurs amis. Et les connaissances de leurs connaissances. C'est ce qui fait d'eux l'une des meilleures armes contre les bactéries dangereuses	La communauté recensée se compose principalement d'exemplaires de la flore vaginale et intestinale maternelle, de germes cutanés et d'une sélection de ce que l'hôpital a en ce moment à proposer. C'est un très bon mélange pour commencer.
Le sucre ménager, par exemple, n'est pas un prébiotique parce que les bactéries des caries l'apprécient aussi. Les prébiotiques ne sont que peu, voire pas du tout assimilés par les mauvaises bactéries, de sorte qu'elles ne peuvent pas s'en servir pour fabriquer des armes contre nous	Quand on parle de cette famille, on ne peut pas dire que les grands esprits (des scientifiques) se rencontrent : c'est la pomme de discorde. Les uns – ceux-là mêmes qui ont vérifié l'existence des entérotypes – n'ont pu trouver que les familles Prevotella et Bacteroides.
Les travaux visant à élaborer une carte des bactéries n'ont commencé qu'en 2007. Avec un coton-tige, on pratique des prélèvements sur un très très grand nombre de participants – en trois endroits de la bouche, sous les aisselles, sur le front, etc.	Quelqu'un a laissé le lait traîner dehors, des bactéries (venues directement de la vache ou bien présentes dans l'air au moment de la traite) ont pénétré dans le récipient, le lait s'est épaissi et, surprise ! un nouvel aliment était né.
La sérotonine, c'est ce neurotransmetteur qu'on surnomme aussi l'"hormone du bonheur" parce qu'une carence peut engendrer des dépressions. Non décelée, une malabsorption du fructose qui dure depuis plusieurs années peut donc tout à fait être la cause d'humeurs dépressives	Si vous publiez des photos de votre dîner sur Facebook et vous étonnez qu'aucun de vos amis ne commente votre chef-d'œuvre, sachez-le : vous vous êtes tout simplement trompé de public. Sur un Facebook microbien, un million d'abonnés applaudirait votre cliché à tout rompre ou frissonnerait de peur.
Le nerf vague est la voie de communication la plus importante et la plus rapide entre l'intestin et le cerveau. Il traverse le diaphragme, passe par le médiastin (la région entre les poumons qui contient notamment le cœur), longe l'œsophage, monte dans le cou et arrive au cerveau	Pour nos enzymes digestives, l'acide lactique lévogyre, c'est un peu comme mettre le pied droit dans la chaussure gauche : inconfortable. Au supermarché, mieux vaut donc choisir les yaourts qui contiennent des bifidobactéries, capables de produire l'acide lactique dextrogyre.

Table B.22 Keywords and duration for list 11 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
découper	bactéries	rainures	personne	lardons	chair	effet	coagulation	yaourt	11
œufs	ferme	poulet	mères	litre	route	dilution	humide	malades	12
chaînes	sucré	pain	menthe	draps	douche	danger	crachat	propreté	13
microbes	sortir	laboratoire	intelligent	combattre	mauvaise	neutraliser	œufs	maladie	13
Tokyo	Châtillon-sur-Seine	renseignements	enceinte	enfant	rapide	africains	éloignement	germes	14
bénéfique	insecte	gratter	mère	voiture	empire	supermarché	tofu	protéine	14
allergies	fragmenter	aminés	cage	maisonnette	chat	bile	affronts	émus	15
tuer	famille	armes	flore	hôpital	mélange	cracher	serviette	distance	15
caries	bactéries	armes	esprits	discorde	existence	remontées	oesophages	brûlure	16
carte	prélèvements	aisselles	dehors	vache	épaissi	Irlande	intestins	phénomène	16
neurotransmetteur	dépressions	fructose	dîner	public	cliché	taille	installer	fixation	17
communication	diaphragme	cou	chaussure	supermarché	yaourt	rhume	cholestérol	mollet	18

Table B.23 Target and masker stories for list 12 of the Long-SWoRD test v1

Target story	Masker story
Notre organe olfactif est un goûteur chevronné. Plus la petite cuillère chargée d'une première bouchée de gâteau se rapproche de la bouche, plus il y a de molécules qui s'en détachent et affluent vers les narines	Par rapport à d'autres régions, c'est aussi dans ces pays qu'on recense un nombre bien plus grand d'infections impossibles à traiter chez l'homme. En France, il y a des règles, mais il faut avouer qu'elles sont très floues
Il s'envola pour l'Amérique du Sud, construisit sur place un radeau selon les techniques de construction de l'époque, emporta avec lui des noix de coco et quelques boîtes d'ananas au sirop et mit le cap sur la Polynésie.	La biotine ne nous sert pas seulement à avoir "une peau rayonnante, des cheveux brillants et des ongles renforcés", comme le vantent certains emballages de comprimés vendus en parapharmacie
Une partie non négligeable de notre poids est induite par les atomes inspirés – et non par le saucisson-beurre. D'ailleurs, les plantes tirent la plus grande part de leur poids non pas de la terre, mais de l'air !	La variante agressive possède deux spécificités : la première, c'est la protéine "cag A", une sorte de minuscule seringue avec laquelle la bactérie peut injecter certaines substances dans nos cellules.
Étrangement, l'odeur est la seule impression sensorielle dont on ne peut pas rêver. Les rêves sont toujours sans odeur. Or, une odeur est capable de faire naître un sentiment d'attirance	le reflux gastro-œsophagien et les remontées acides sont et restent de petites erreurs de parcours désagréables, mais anodines. Ce n'est pas plus grave que de rater une marche : on se relève, on remet de l'ordre dans sa tenue
Aujourd'hui, qui veut explorer de nouveaux continents et rencontrer de nouveaux peuples doit partir à la découverte du "petit monde" qui se trouve en nous. L'intestin est le continent le plus fascinant de ce monde-là	C'est là qu'interviennent les mucines. Ces protéines visqueuses contenues dans notre salive offrent à chacun de nous quelques heures de grand spectacle – quand, enfant, nous nous apercevons que nous pouvons faire des bulles avec notre salive.
Bientôt, les antibiotiques pourraient être remplacés par un extrait concentré de brocoli – le sulforaphane. Il s'agit d'une substance capable de bloquer l'enzyme qui permet à Helicobacter de neutraliser l'acidité gastrique	Nul besoin de voir grand, commençons par de petites choses comme nos repas quotidiens, en suivant là aussi cette règle : pas de stress, pas de tensions. Les repas devraient être des zones de calme, sans dispute
Nous autres humains sommes très fiers de la complexité de notre cerveau. Réfléchir sur des lois fondamentales, des questions philosophiques et religieuses ou encore des problèmes de physique est une performance de tout premier plan	Un morceau d'entrecôte peut par exemple se balancer six heures dans notre petit hamac avant d'être livré intégralement à l'intestin grêle. Pas étonnant, donc, que nous ayons une terrible envie de dessert après avoir mangé de la viande ou des beignets bien gras
Il y a une maladie dont le symptôme principal consiste en une perception sensorielle erronée : c'est la schizophrénie. Les patients ont par exemple l'impression que des fourmis leur grimpent sur le dos alors qu'il n'y a pas la moindre bestiole alentour	La fermentation produit souvent des acides, qui donnent par exemple au yaourt ou aux légumes fermentés un goût plus acide que l'aliment de départ. L'acidité et les bonnes bactéries protègent la nourriture des mauvaises bactéries
95 % de la sérotonine que nous produisons nous- mêmes est fabriquée... où ? Dans les cellules de l'intestin. Elle est là pour prêter main-forte aux nerfs qui président aux mouvements des muscles et sert aussi de molécule transductrice essentielle.	l'émail de nos dents se défend bien lui aussi, puisque c'est le matériau le plus dur que nous soyons capables de fabriquer. Un record qui a sa raison d'être : avec notre mâchoire, nous exerçons sur une molaire une pression qui peut aller jusqu'à 80 kilos,
Les bactéries sont riches en protéines : d'un point de vue nutritionnel, on peut donc les considérer comme de petits biftecks. Quand elles ont fait leur temps dans l'estomac bovin, elles prennent le toboggan vers les étages inférieurs où elles sont alors digérées.	bien-être, joie, satisfaction – tout relève pour nous de la tête. Et quand nous n'avons pas confiance en nous, quand nous sommes inquiets ou dépressifs, nous avons honte de loger à l'étage supérieur un ordinateur défaillant.
La recherche sur les bactéries permet en outre d'avancer que le stress est antihygiénique. Les bactéries qui survivent dans l'intestin quand les conditions vitales ont changé ne sont en effet pas les mêmes que quand on se la coule douce. Le stress, pour ainsi dire, influe sur la météo intestinale	Et les bébêtes sont patientes : elles peuvent attendre jusqu'à cinq ans avant d'emménager chez leur nouvel hôte. Les propriétaires de chat ne sont donc pas les seuls visés. Les chats et les autres animaux hôtes prennent l'air dans les jardins, ils se promènent dans les champs plantés de légumes ou sont parfois tués.
Les effets secondaires d'antidépresseurs courants comme le Prozac nous apportent par ailleurs des renseignements importants sur la sérotonine. Un patient traité sur quatre fait face aux effets typiques que sont la nausée, la diarrhée et, en cas de prise prolongée, le ralentissement du transit.	après avoir fait l'expérience d'un monde complètement étranger, de baisers par milliers, de promenades en forêt ou de jeux dans le jardin, de poils de chien ou de chat, de rhumes à répétition et d'un tas d'enfants inconnus à l'école, notre système immunitaire a fini ses études

Table B.24 Keywords and duration for list 12 of the Long-SWoRD test v1

Target 1	Target 2	Target 3	Mask 1	Mask 2	Mask 3	Extraneous 1	Extraneous 2	Extraneous 3	Sec.
chevronné	gâteau	molécules	pays	infections	France	film	neutralisées	salive	12
Amérique	construction	ananas	peau	ongles	comprimés	velouté	hérissées	cerf	12
atomes	saucisson	terre	spécificité	protéine	substance	remèdes	ablation	réponse	13
sensorielle	rêves	sentiment	remontées	désagréable	marche	résistance	pompes	cave	13
peuples	monde	fascinant	protéines	spectacle	bulles	mitonnée	chaîne	école	14
remplacés	brocoli	enzyme	commençons	règle	calme	billard	jouer	profonds	14
complexité	philosophiques	physique	hamac	dessert	viande	enzyme	envahisseurs	somnolence	15
schizophrénie	fourmis	bestiole	yaourt	aliments	nourriture	porteurs	probabilité	vasculaire	15
sérotonine	nerfs	molécule	matériaux	record	pression	régulent	marche	obstacle	16
protéines	biftecks	toboggan	tête	inquiets	ordinateur	grossesse	répartition	regardés	16
survivent	vitales	météo	patientes	emménager	jardins	catastrophique	cancer	humanité	17
Prozac	sérotonine	nausée	baisers	jardin	rhumes	ablation	durable	cohabitaient	18

Appendix C

Long - SWoRD v2 and rhythmic priming experiment

C.1 Selection of the speech material

The material of the second version of Long-SWoRD test was used in a rhythmic priming experiment run by Anna Fiveash from the Center of Research in Neurosciences of Lyon (CRLN). The task was to decide if a keyword belonged to the story that the participants listen to in an isolated condition.

470 stories were selected from the 700 available based on 5 criteria. The stories had to be (1) interesting, (2) intelligible, (3) contain two or three sentences, (4) last between 11 and 16 seconds and (5) contain no questions. The language frequency of all the keywords was also estimated using the Lexique database (New et al., 2001) in the same way as for the material in version 1 of the test. Stories that had keywords not in the Lexique database were removed.Then, 96 stories were matched by 4 based on their duration, their number of sentences and their number of syllables.

The keywords and their distractors were already selected. The keywords were not too rare, not too frequent and only appear once in each story. The distractor

keywords were matched with two online synonyms databases (CRISCO [1] and CNRTL [2])

C.2 Final version of the material

Table C.1 4 stories for set 1

Story
"Près d'un tiers des oiseaux d'Amérique du Nord sont en voie d'extinction. Selon un rapport alarmant, un virus issu des élevages intensifs de volailles serait responsable de cette catastrophe. Il rendrait les femelles stériles."
'Suite au succès phénoménal de certaines séries américaines, les producteurs ont trouvé une nouvelle source financière ! Les sites sur lesquels les tournages se sont déroulés sont maintenant accessibles. Ces lieux pourront réjouir tous les fans !'
"Dans quelques dizaines d'années, l'espace deviendra peut-être une destination ordinaire. Coûteux et physiquement contraignant, le tourisme spatial pourrait quand même voir le jour. Le coût du voyage sera le même qu'une belle voiture."
"Notre maison abriterait près de deux cent mille types de champignons et bactéries. Inutile de tout nettoyer, ils sont inoffensifs. Des variations existent entre les foyers et sont surtout liés à la présence d'animaux domestiques."

Table C.2 Keywords and duration for list 1

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
extinction	virus	volailles	extinction	disparition	12
séries	sites	réjouir	réjouir	enchanter	12
espace	contraignant	voyage	espace	cosmos	12
types	nettoyer	variations	types	sortes	12

Table C.3 4 stories for set 2

Story
'Le mythe selon lequel les poissons rouges sont dotés d'une mémoire de trois secondes serait erroné. Il aurait été inventé pour éviter de culpabiliser en les enfermant dans des bocaux minuscules. En réalité, ils auraient une mémoire d'au moins trois mois.'
'Un centenaire résidant avec son épouse en maison de retraite a été signalé disparu pendant plusieurs heures. Sept policiers ont été appelés sur les lieux. Il a finalement été retrouvé, dormant paisiblement dans le lit d'une autre résidente.'
'On peut en observant une statue équestre déterminer les conditions de la mort du cavalier. C'est la position des jambes avant du cheval qui nous le permet. Par exemple, si les deux jambes sont levées, on en déduit que le cavalier est décédé au combat.'
'Toutes les sucreries ne causent pas de caries. En effet, le chocolat noir protège les dents car il est riche en cacao, qui est composé de fluor et de phosphate. Le chocolat au lait, bien plus sucré, a un effet moindre et ne doit pas dispenser du brossage de dents !'

[1] http://crisco.unicaes.fr
[2] www.cnrtl.fr

Table C.4 Keywords and duration for list 2

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
poissons	culpabiliser	réalité	culpabiliser	regretter	12
épouse	policier	lit	lit	couchette	12
statue	cheval	levées	statue	sculpture	12
caries	cacao	lait	caries	lésions	12

Table C.5 4 stories for set 3

Story
'Les traumatismes crâniens peuvent avoir de drôles de conséquences. Un jeune américain tombé dans le coma suite à un choc s'est réveillé en parlant couramment espagnol alors qu'il n'en connaissait que les bases. Sa langue natale lui est revenue peu à peu.'
'Vous avez plus de chances de tomber malade après une averse. En effet, les gouttes de pluie tombant dans les flaques produisent des bulles qui explosent en projetant dans l'air des bactéries présentes habituellement sur le sol. Ces bactéries restent alors en suspension dans l'air.'
'Il est possible de se muscler par la pensée. En effet, les médecins conseillent aux personnes blessées, fracture par exemple, de s'imaginer en train de stimuler les parties du corps atteintes. Ceci a pour but d'amoindrir drastiquement la perte de leurs facultés.'
'Le terme « atterrir » est utilisé pour tout atterrissage sur une surface solide. Le mot terre fait ici référence à la terre ferme et non à la planète Terre. On peut donc atterrir sur Mars et même par extension atterrir sur un lac de méthane de Titan.'

Table C.6 Keywords and duration for list 3

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
crâniens	choc	natale	choc	collision	13
malade	exposent	suspension	explosent	éclatent	13
pensée	stimuler	amoindrir	pensée	réflexion	13
surface	planète	lac	lac	mer	13

Table C.7 4 stories for set 4

Story
"Le café a des vertus qui pourraient donner envie d'en boire régulièrement. Il permettrait de lutter contre l'apparition de cancers de la peau. De plus c'est un remède utile contre le stress ou encore un bon moyen d'empêcher son foie de stocker la graisse."
"Pour mieux dormir ce soir, n'essayez pas de trouver le sommeil à tout prix ! Cela peut paraître étonnant, mais la meilleure façon de s'assoupir est parfois de ne pas le chercher. Les insomniaques ont plus de chance de s'endormir en essayant de rester éveillés."
'Un hôtelier asiatique a installé dans la cour de son établissement deux containers en guise de chambres. Ils disposent naturellement de tout le confort des autres suites. Voilà une façon originale d'attirer de nouveaux vacanciers assoiffés d'exotisme.'
'Selon une étude, les étudiants peuvent lutter contre l'anxiété en écrivant leurs inquiétudes avant le début d'un examen. Cela permettrait de décharger les angoisses et de libérer la puissance intellectuelle. Cela leur permettrait de mieux se concentrer.'

Table C.8 Keywords and duration for list 4

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
boire	cancers	remède	boire	s'abreuver	13
trouver	assoupir	chance	assoupir	somnoler	13
cour	confort	attirer	attirer	conquérir	13
étudiants	examen	intellectuelle	étudiants	élèves	13

Table C.9 4 stories for set 5

Story
'Boire son Coca-Cola dans un sac plastique peut paraître étrange, mais c'est une pratique répandue en Amérique Centrale. Cette habitude permet aux consommateurs de payer leur boisson moins cher. En effet, cela leur permet d'acheter la boisson sans bouteille ou cannette.'
'Au football américain, la plupart des joueurs ont de grandes traces noires placées sous les yeux. Leur rôle premier est non pas de leur donner un côté guerrier, mais de diminuer la réflexion de la lumière des projecteurs du stade ou du soleil. Cela évite aux joueurs d'être aveuglés.'
'La fonte n'est pas plus lourde que l'acier ou le fer, c'est même l'inverse. Cependant, elle est souvent utilisée pour la production d'objets massifs comme des radiateurs, poêles ou enclumes car elle ne coûte pas cher. Cela explique pourquoi elle est souvent associée aux objets lourds. '
'Il n'existe pas de trous d'air au sens propre en avion. Les chutes soudaines ou les turbulences sont dues aux remous d'air, c'est à dire à des courants désordonnés, parfois descendants. Dans certains de ces remous, l'avion peut perdre des dizaines de mètres subitement, sans danger.'

Table C.10 Keywords and duration for list 5

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
plastique	payer	cannette	payer	acheter	13
traces	guerrier	évite	traces	marques	13
fer	massifs	associée	massifs	imposants	13
trous	courants	perdre	courants	vents	13

Table C.11 4 stories for set 6

Story
'Il existe un syndrome de Paris. Il s'agit d'une dépression vécue par certains touristes notamment asiatiques déçus de la ville, par rapport à son image idyllique véhiculée par les films. A cela s'ajoutent le fait qu'ils sont victimes de nombreux vols.'
'Les hublots d'avion comportent tous un petit trou dans leur partie basse. Cela permet de rediriger la pression de l'appareil directement vers la vitre extérieure. C'est donc une mesure de sécurité indispensable pour pallier à toute destruction de cette dernière.'
'Lorsqu'on lit, notre cerveau traite l'information de plusieurs manières en même temps. Il traite les lettres une par une tout en analysant le mot dans son ensemble. Ainsi, tant que la première et la dernière lettre des mots restent en place, notre cerveau s'organise pour reconstruire le tout.'
'Lorsque vous faites du sport et brûlez vos graisses, celles-ci ne partent pas dans la transpiration mais majoritairement dans la respiration. Le mécanisme qui brûle les graisses est chimiquement complexe. Il les transforme en grande partie en dioxyde de carbone que l'on expire.'

Table C.12 Keywords and duration for list 6

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
dépression	asiatiques	victimes	dépression	déception	13
trou	pression	destruction	destruction	démolition	13
information	mot	reconstruire	information	contenu	13
sport	mécanisme	dioxyde	sport	activité	13

Table C.13 4 stories for set 7

Story
'Mettre de l'ordre dans sa maison est aussi une manière de ranger son esprit. C'est ce que Marie Kondo met à l'honneur dans son livre « la magie du rangement ». Sa méthode est l'art de l'organisation, elle permet de faire le tri et de réorganiser de façon spectaculaire sa maison.'
'Encore un peu endormie, Marie se rend dans sa salle de bain afin de faire sa toilette avant de se rendre au travail. Elle commence par se passer un peu d'eau sur le visage. Au moment de prendre sa serviette pour s'essuyer, elle y aperçoit une grosse araignée noire, ce qui la fait bondir de frayeur.'
'En moins huit avant Jésus Christ, l'empereur Auguste décida de rajouter un jour au mois d'août qui porte son nom, pour en avoir trente-et-un. Soit autant que le mois de juillet qui lui était appelé ainsi en l'honneur de Jules César. Ce jour a été substitué au mois de février.'
'Le fameux point de côté serait dû à une crampe musculaire du diaphragme. En effet, pendant un effort physique, la respiration s'accélère ce qui peut engendrer la douleur familière. Une autre hypothèse repose sur une trop grande accumulation de sang dans le foie ou la rate pendant l'effort.'

Table C.14 Keywords and duration for list 7

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
manière	livre	réorganiser	manière	façon	13
toilette	visage	araignée	toilette	douche	13
rajouter	juillet	substituer	substitué	retiré	14
crampe	douleur	sang	sang	hémoglobine	13

Table C.15 4 stories for set 8

Story
"Une grande marque de joaillerie parisienne recrute pour sa prochaine campagne publicitaire. Pour mettre en avant leur bijou, l'entreprise recherche une voix très spéciale. De grandes stars de la chanson française ont déjà passé le casting sans avoir été sélectionnées."
"Sur les trottoirs argentins, des frigos remplis de nourriture attendent d'être ouverts. Ces réfrigérateurs en libre-service sont apparus pour éviter le gaspillage alimentaire. De plus, il permet de faire face à l'explosion du chômage et de l'inflation."
'La raie pastenague est une espèce de raie venimeuse vivant dans les eaux tropicales. Son dard est si fin qu'in simple contact suffit pour le faire pénétrer sous la peau et libérer le venin. Les symptômes arrivent de manière crescendo : paralysie, fièvre, spasmes.'
'On entend souvent dire que le muscle le plus puissant du corps humain par rapport à sa taille est la langue. Ceci est faux car la langue est un organe composé de dix-sept muscles différents. Le muscle le plus puissant du corps humain, qui est aussi le plus grand, est le grand fessier. '

Table C.16 Keywords and duration for list 8

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
campagne	bijou	chanson	campagne	stratégie	13
remplis	éviter	chômage	chômage	inactivité	14
venimeuse	contact	symptôme	venimeuse	envenimé	13
taille	faux	dix-sept	faux	erroné	13

Table C.17 4 stories for set 9

Story
'Les touristes seront étonnés de savoir que la majorité des plages françaises sont artificielles. Pour répondre à une demande économique, des plages ont été créées durant les années soixante-dix. Tout le littoral est régulièrement réapprovisionné en sable et galets.'
"A l'origine, l'astrologie était utilisée par les dirigeants de nombreuses régions. Dans l'Antiquité, elle servait à connaître le moment le plus favorable pour une récolte, construire un monument ou attaquer des ennemis. Rares étaient ceux qui la remettaient en cause."
'Lorsque vous faites cuire des steak-hachés frais ou surgelés, placez en premier côté poêle le côté lisse du steak. En effet, les rayures ont été faites exprès afin d'évacuer facilement le sang et la graisse lors de la cuisson. Celle-ci sera donc meilleure en terminant par la face rayée.'
'En Suisse, la production, l'abattage et la vente de chiens et de chats sont interdits. Cependant, leur consommation est autorisée. Il n'y a pas de statistiques fiables sur le pourcentage de la population qui en consomme, mais cela resterait faible et limité à certains cantons.'

Table C.18 Keywords and duration for list 9

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
majorité	répondre	littoral	majorité	plupart	14
dirigeants	récolte	attaquer	récolte	cueillette	14
cuire	évacuer	meilleure	meilleure	préférable	14
abattage	autorisée	faible	faible	moindre	14

Table C.19 4 stories for set 10

Story
'Il est fortement déconseillé de donner du miel aux enfants de moins d'un an. En effet, les abeilles peuvent transporter en même temps que le pollen des spores de la bactérie responsable du botulisme. Avant un an, le système immunitaire de l'enfant est trop immature pour lutter contre cette bactérie.'
'Contrairement aux croyances, il est impossible d'allumer un feu en frottant deux silex l'un contre l'autre. En effet, les étincelles produites sont trop volatiles pour offrir une chaleur suffisante à l'inflammation d'herbes sèches. Seuls les minerais contenant du sulfate de fer peuvent allumer un feu.'
'Il n'est pas possible de briser un verre avec une voix humaine. Pour le faire, il faudrait qu'un chanteur puisse tenir la note correspondant exactement à la fréquence de résonance du verre. Même les plus grands artistes lyriques n'en sont pas capables, leur voix n'étant jamais stabilisée sur une fréquence précise.'
'La théorie de l'esprit n'est pas réservée aux humains, les corbeaux par exemple, en seraient dotés. Une équipe autrichienne a étudié leur comportement et a mis en évidence que ces oiseaux comprenaient qu'ils pouvaient être observés. Ils pourraient même agir en conséquence, souvent de façon sournoise.'

Table C.20 Keywords and duration for list 10

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
déconseillé	pollen	immature	déconseillé	contre-indiqué	14
croyances	volatiles	fer	volatiles	éphémère	14
briser	note	stabilisée	note	son	14
esprit	comportement	agir	esprit	conscience	14

Table C.21 4 stories for set 11

Story
"Une britannique de cinquante-trois ans, dans le coma depuis plusieurs années, est revenue soudainement à la vie. L'écoute des premières notes de sa chanson préférée lui a permis de sortir de cet état. Cet événement offre de nouvelles perspectives sur le niveau de conscience."
'Les arcs-en-ciel ne sont pas toujours colorés. Très rarement, il peut se produire un phénomène durant lequel de fines gouttes d'eau sont traversées par la lumière dans la brume et le brouillard. La lumière n'est alors pas décomposée et l'arc-en-ciel apparaît de couleur blanche.'
'Les noms des rues sont nés durant le Moyen-âge d'une manière très simple. Les rues étaient nommées en fonction de leur situation géographique et des endroits qu'elles desservaient. Plus tard, ont été intégrés des valeurs, batailles, personnages artistiques, historiques et politiques notamment.'
'Si un jour il commence à pleuvoir des poissons, n'ayez pas peur ! Les pluies de poissons sont fréquentes à travers le monde, et seraient liées à des phénomènes météorologiques. En effet, les fortes tempêtes peuvent emporter les petits poissons et ensuite les faire retomber sous forme de pluie.'

Table C.22 Keywords and duration for list 11

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
coma	chanson	événement	chanson	refrain	14
colorés	gouttes	décomposée	colorés	teinté	14
nés	situation	historiques	situation	emplacement	14
pleuvoir	phénomène	retomber	phénomènes	condition	14

Table C.23 4 stories for set 12

Story
'Le syndrome de l'huître est un phénomène connu exploité par les techniques de marketing. Il consiste en la satisfaction que retire un consommateur de la difficulté qu'il aura à ouvrir un emballage. C'est la raison pour laquelle certains produits sont difficiles à déballer.'
'Des rosiers sont souvent plantés aux extrémités des rangs de vignes. Cela est dû au fait que la vigne et les rosiers sont sensibles aux mêmes maladies. Le rosier étant toujours atteint plus tôt, il permet d'alerter le vigneron qui peut alors intervenir et traiter la vigne en prévention.'
'Un chien avait l'habitude d'accompagner son maître à la gare pour qu'il aille travailler, et revenait le soir à la même place pour attendre son retour. Lorsque son maître décéda, il continua à revenir tous les soirs à la gare, jusqu'à sa mort. Une statue fut érigée en son honneur.'
'Après un déjeuner en famille, Régis sortit de la maison pour se rendre à son rendez-vous. Quelques minutes plus tard, il demanda à son fils de le rejoindre dehors. Celui-ci découvrit alors avec stupeur que son père n'ayant pas fait attention, avait embouti sa voiture neuve.'

Table C.24 Keywords and duration for list 12

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
huître	satisfaction	produits	satisfaction	bonheur	14
rangs	maladies	traiter	traiter	soigner	14
travailler	revenir	statue	travailler	exercer	14
maison	dehors	embouti	embouti	percuté	14

Table C.25 4 stories for set 13

Story
'Les sodas ne contiennent que des calories « vides ». En effet, ils ne comportent pas de nutriments essentiels : pas de vitamines, pas de minéraux, pas d'antioxydants, ni de fibres. Ils n'apportent rien à votre alimentation à part de grosses quantités de sucre et de calories dont vous pouvez vous passer.'
'Une femme, aux Etats-Unis, a été arrêtée en état d'ébriété sur la route. Cependant, elle ne se trouvait pas au volant d'une voiture mais bien sur un cheval ! Selon les autorités, elle a mis en danger sa vie, celle du cheval et des usagers de la route et a été placée en détention.'
'Vous vous êtes peut-être déjà dit que lorsque vous lisez ou pensez, vous entendez une petite voix intérieure. Des chercheurs lyonnais ont réussi à prouver que cette voix existait bel et bien, et qu'elle était créée par notre cerveau. Son origine viendrait de l'apprentissage de la lecture à haute voix.'
'Nos expressions faciales seraient innées et non apprises. En effet, une étude a été menée sur les réactions d'athlètes aux Jeux Olympiques et Paralympiques. Elle a démontré que des personnes aveugles de naissance utilisaient les mêmes expressions faciales lors du même type d'émotions que les voyants.'

Table C.26 Keywords and duration for list 13

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
vides	nutriments	sucre	vides	absente	14
arrêtée	voiture	danger	voiture	automobile	14
intérieure	prouver	apprentissage	apprentissage	acquisition	14
innées	athtlètes	expressions	athlètes	sportifs	14

Table C.27 4 stories for set 14

Story
'La « lévitation acoustique » était déjà connue, mais seulement limitée à un axe fixe. C'est-à-dire que les objets ne pouvaient pas se déplacer via cette technique. Dernièrement, des scientifiques japonais ont réussi à se faire déplacer de petits objets à l'aide de simples ondes sonores.'
'Le lait est une émulsion naturelle, c'est-à-dire qu'il est composé d'un liquide dispersé dans un autre avec lequel il ne se mélange pas. De microscopiques gouttelettes de matière grasse sont en suspension dans de l'eau. La couleur blanche est due aux protéines et minéraux qu'il contient.'
' Le Japon détient le record du plus grand nombre de centenaires jamais enregistré à l'échelle d'un pays. A ce jour, près de soixante mille centenaires vivent joyeusement au Japon. Cela pose un problème au gouvernement, car le budget cadeau pour ceux qui fêtent leur centenaire devient trop important.'
'Le roquefort n'est pas seulement bon en bouche, il a aussi un pouvoir guérisseur. Il contient naturellement de la pénicilline qui agit comme antibiotique et empêche la formation de la paroi de la bactérie. En cas d'infection bactérienne bénigne, pensez d'abord au roquefort !'

Table C.28 Keywords and duration for list 14

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
limitée	technique	japonais	technique	approche	13
liquide	suspension	blanche	liquide	fluide	13
record	vivent	budget	record	exploit	13
bouche	formation	infection	infection	contamination	13

Table C.29 4 stories for set 15

Story
"Tenir une conversation à l'étranger sans redouter la barrière de la langue sera désormais possible. Un écouteur qui traduit presque instantanément les discussions pourrait être mis sur le marché dès fin deux mille seize. Pour le moment quatre langues sont proposées."
"Selon un magazine américain, le meilleur croissant du monde n'est pas français ! Ils ont été séduits par les préparations d'une boulangerie australienne. Celle-ci propose, en plus des croissants traditionnels, des versions salées, fourrées au fromage ou au piment."
"Pour se muscler, il suffit de penser ! Il peut suffire, pour renforcer ses muscles, qui sont dans un plâtre, par exemple, d'imaginer qu'on les contracte. Ceci nécessite un exercice mental soutenu et très régulier, mais démontre bien la force de la pensée sur notre corps."
'Paradoxalement, les forces spéciales américaines ont perdu plus d'hommes sur leur territoire qu'au combat. Cela est dû au fait que beaucoup d'hommes se suicident à leur retour au pays. Ce phénomène est la cause du stress post-traumatique subi lors des conflits armés.'

Table C.30 Keywords and duration for list 15

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
étranger	traduit	marché	marché	commerce	14
américain	préparations	salées	salées	épicées	14
renforcer	exercice	démontre	exercice	entraînement	14
spéciales	retour	stress	retour	rapatriement	14

Table C.31 4 stories for set 16

Story
'Notre consommation de biens matériels est régie par l'obsolescence programmée. Ce principe consiste à réduire sciemment la durée de vie des objets. Une fois hors service, le consommateur n'a pas d'autres choix que de se réapprovisionner.'
'Une petite fille atteinte d'une maladie très rare avait besoin d'une greffe de rein. Mais elle ne trouvait pas de donneur. Alors, sa maîtresse entreprit les démarches afin de vérifier sa compatibilité et pu lui faire don d'un de ses reins.'
'Après avoir épluché des oignons, l'odeur sur les doigts est difficile à faire partir. Un simple lavage des mains ne suffit pas. Il suffit de passer la lame d'un couteau sur ses doigts, ou tout autre objet métallique pour s'en débarrasser.'
'La police néerlandaise a mis au point un procédé étonnant pour faire face à une éventuelle menace de drones survolant des sites sensibles. Ils ont dressé des aigles. Les rapaces ont appris à neutraliser et rapporter les drones aux forces de l'ordre.'

Table C.32 Keywords and duration for list 16

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
matériels	réduire	choix	choix	alternative	11
maladie	maîtresse	compatibilité	compatibilité	correspondance	11
oignons	lavage	métallique	métallique	ferrailleux	11
procédé	aigles	neutraliser	neutraliser	combattre	11

Table C.33 4 stories for set 17

Story
'Le gaspillage alimentaire pourrait être évité. Certains plats peuvent être utilisés des jours voire des semaines après la date limite de consommation. Il est possible de manger du miel plusieurs années après ouverture.'
'Les épinards contiennent en réalité peu de fer. L'idée qu'ils en contiennent beaucoup provient d'une erreur de virgule. Les aliments les plus riches en fer sont le boudin, les viandes rouges en général, mais aussi le thym.'
'Un jeune entrepreneur canadien s'est lancé dans une activité très originale. Il recycle les vieux bijoux de ses clients. Il se charge de fondre les métaux afin de créer par la suite un nouveau modèle pour sa marque.'
'Bien qu'il existe différents types de thés, leurs origines restent néanmoins communes. Les feuilles proviennent toutes du même arbuste. Mais la main humaine, à travers des processus de fabrication plus ou moins complexes, crée cette différenciation.'

Table C.34 Keywords and duration for list 17

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
machines	supermarchés	accueillir	limite	maximum	11
réalité	virgule	rouges	virgule	ponctuation	11
activité	fondre	modèle	modèle	prototype	11
thés	arbuste	fabrication	arbuste	arbre	11

Table C.35 4 stories for set 18

Story
'Les principaux prédateurs des pucerons sont les coccinelles mais aussi les larves d'une certaine famille de mouche. Les plantes, pour se protéger des pucerons, produisent une substance qui attire les coccinelles. Ces dernières, en arrivant sur l'arbre, se régalent de cette colonie parasite.'
'Ils naviguaient depuis plusieurs jours déjà, quand ils aperçurent une forme au loin. D'abord floue, elle gagna en précision quand ils se rapprochèrent. Quand le capitaine entendit les cris de joie de l'équipage, il sut qu'ils avaient réussi et étaient arrivés à destination.'
'Il existe un syndrome particulier, appelé syndrome de Paris. Il désigne le choc psychologique négatif qu'ont certains voyageurs lorsqu'ils réalisent la différence entre le vrai Paris et leur vision idéalisée. Les Japonais sont les cas les plus souvent répertoriés. '
'Contrairement à ce que l'on croyait depuis toujours, les lions chassent autant que les lionnes mais ils le font de nuit. Ils se cachant dans les zones de végétation dense pour mieux surprendre leurs proies. Une méthode bien moins voyante que celle des lionnes qui elles chassent généralement de jour et dans des zones dégagées.'

Table C.36 Keywords and duration for list 18

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
larves	substance	arbre	arbre	tronc	14
forme	précision	équipage	équipage	marin	14
particulier	voyageurs	cas	voyageurs	visiteurs	13
nuit	dense	voyante	nuit	obscurité	14

Table C.37 4 stories for set 19

Story
'Vous avez peut-être déjà vu sur une carte de restaurant, à côté des andouillettes, l'appellation cinq A. Cela signifie que l'andouillette qui est servie remplit le cahier des charges d'une très sérieuse association. Il s'agit de l'Association Amicale des Amateurs d'Andouillette Authentique.'
'La médaille olympique de bronze a un coût de fabrication étonnant. Contrairement aux médailles d'or et d'argent qui ont un coût de fabrication élevé, elle ne coûte environ que deux euros et quarante centimes à produire. Elle est faite d'un alliage de cuivre, de zinc et d'étain, ce qui explique ce faible coût.'
'Lors d'une activité physique, l'irrigation est modifiée. Par exemple, le volume sanguin de l'appareil digestif et urinaire diminue au profit d'une augmentation du volume sanguin irriguant les muscles. Le cerveau est le seul organe pour lequel le volume distribué ne change pas au cours d'un effort.'
'Un groupe de jeunes filles s'apprêtait à partir en vacances et était bien installé dans l'avion. Les portes venaient de se fermer quand une femme décida de retenir l'avion afin que sa famille, en retard, la rejoigne. Les hôtesses durent appeler les forces de police et l'avion pu décoller avec une heure de retard.'

Table C.38 Keywords and duration for list 19

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
restaurant	cahier	Amateurs	amateurs	gourmets	14
bronze	élevé	alliage	élevé	considérable	14
modifiée	appareil	organe	modifiée	varie	14
vacances	retenir	police	vacances	voyage	14

Table C.39 4 stories for set 20

Story
'La vision aveugle est étudiée par des neurologues. Cette vision est basée sur les perceptions conscientes et inconscientes du cerveau. Cet inconscient neurologique permet aux personnes atteintes de cécité corticale de voir sans en avoir conscience.'
'Un concours du pull de Noël le plus moche a été lancé en octobre 2017. Il invite les participants à ressortir leurs plus étonnants spécimens. Les gagnants remporteront un verre à partager entre amateurs de rennes, Saint Nicolas et autres effigies de Noël !'
'Nous ne soupçonnons pas la quantité de microbes introduite dans nos maisons. Parmi les objets les plus sales, le sac à main est en première place. En effet, souvent posé directement sur le sol, il véhicule un grand nombre de bactéries.'
"Un automobiliste a emprunté l'autoroute en sens inverse. Par chance, le conducteur n'a percuté aucune voiture, il s'est vite mis sur le côté. Il explique avoir voulu faire demi-tour après avoir pris la mauvaise direction."

Table C.40 Keywords and duration for list 20

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
neurologues	cerveau	cécité	neurologues	médecins	13
moche	étonnants	rennes	rennes	cerf	12
microbes	sac	sol	microbes	germes	13
sens	côté	demi-tour	côté	bord	12

Table C.41 4 stories for set 21

Story
'Des chercheurs ont constaté que boire deux pintes de bière réduirait d'un quart la douleur physique. Cet effet est plus puissant que celui du paracétamol. Cela pourrait expliquer l'abus d'alcool chez les personnes souffrant de douleurs persistantes.'
'Une étude australienne a révélé qu'une voiture noire a plus de risque d'avoir un accident en plein jour qu'une voiture blanche. Les couleurs sombres ont un indice de visibilité moins élevé. Mais ces résultats ne sont pas valables la nuit.'
'Le rouge à lèvre, ce produit cosmétique dangereux. Le parlement britannique a fait voter une loi pour l'interdire en accusant les femmes de s'en servir comme moyen de séduction. Ces dernières pouvaient être reconnues coupables de sorcellerie.'
'En Chine, il existe trois restaurants où les serveurs et les cuisiniers sont des robots. Ces derniers ne tombent pas malades, et ne réclament pas de vacances ni de jours de congés. Leurs principaux avantages sont donc leur efficacité et leur rentabilité.'

Table C.42 Keywords and duration for list 21

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
pintes	puissant	alcool	puissant	efficace	11
voiture	accident	sombres	accident	accrochage	11
cosmétique	interdire	séduction	séduction	attraction	12
restaurants	malades	avantages	avantages	intérêts	12

Table C.43 4 stories for set 22

Story
'L'estomac est un organe dont la taille peut énormément varier. Lorsqu'il est vide, il peut être plus petit que le poing, mais il peut multiplier son volume par vingt à l'occasion d'un repas. Pour cela, la paroi gastrique devient de plus en plus lisse, à mesure que l'organe se remplit.'
'La résistance des peuples de l'Himalaya et notamment des Tibétains à la raréfaction de l'air résulte de gènes spéciaux dont ils disposent. Ceux-ci leur permettent d'utiliser moins d'oxygène pour la respiration. Cela est utile en altitude, où l'oxygène est plus rare.'
'Lorsque les films étaient encore tournés en noir et blanc, les acteurs avaient des astuces de maquillage pour mieux apparaître à l'écran. Les actrices par exemple utilisaient du vert autour des yeux. Cela donnait à l'écran un rendu gris pâle, donnant l'impression d'avoir des yeux plus petits.'
'Ce n'est pas parce qu'elle est bio que l'agriculture n'utilise pas de pesticides. Elle doit utiliser des pesticides naturels, ce qui ne veut pas forcément dire qu'ils sont sans danger. Le sulfate de cuivre, par exemple, est un fongicide naturel toxique pour les humains et les animaux.'

Table C.44 Keywords and duration for list 22

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
taille	multiplier	lisse	taille	volume	13
peuples	respiration	altitude	altitude	montagne	13
astuces	vert	rendu	rendu	effet	13
agriculture	danger	toxique	toxique	nocif	13

Table C.45 4 stories for set 23

Story
'En cette chaude journée d'été, Xavier ne put résister à la tentation de se dévêtir pour plonger dans les eaux fraîches de la rivière. Après avoir nagé quelques mètres, il sentit un étrange frôlement le long de sa jambe. Il vit alors avec effroi qu'un banc de piranhas s'approchait de lui.'
'A force d'utiliser des traitements de texte sur ordinateur, l'écriture manuelle laisse parfois sentir le manque du correcteur orthographique. Deux allemands ont donc souhaité mettre un terme à ce genre de désagrément. Ils ont créé un stylo qui identifie les fautes d'orthographe et les lettres mal formées.'
'La télévision exerce un tel effet sur notre cerveau que même nos rêves sont affectés. En fait, depuis qu'on a la télévision en couleur, la majorité de nos rêves le sont aussi. En comparaison, les gens ayant grandi avec la télévision en noir et blanc rêveraient plus souvent en noir et blanc.'
'Les fourmis de feu, originaires d'Amérique du Sud, possèdent un remarquable instinct de survie. Confrontées à un risque de noyade, elles font preuve d'un élan de solidarité et s'accrochent les unes aux autres pour former un radeau. Cela permet à leur espèce de survivre aux moussons et aux inondations.'

Table C.46 Keywords and duration for list 23

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
tentation	étrange	effroi	étrange	curieux	14
ordinateur	désagrément	fautes	désagrément	inconvénient	14
effet	couleurs	grandi	effet	impact	14
instinct	solidarité	espèce	solidarité	entraide	14

Table C.47 4 stories for set 24

Story
"L'arc-en-ciel ne contiendrait pas sept couleurs mais beaucoup plus. Cette croyance perdure depuis le début du dix-huitième siècle. En fait, il présente un nombre infini de couleurs qui échappent à notre regard."
"Pas de baisers d'adieu ! Il est interdit, depuis deux mille neuf, de s'embrasser dans les véhicules stationnant devant une gare anglaise. Les longs au revoir des amoureux provoquaient trop d'embouteillages."
'Une étude récente a confirmé que nous ne perdons pas beaucoup de chaleur par la tête. Toute partie du corps qui n'est pas couverte perd de la chaleur. Ainsi, s'il fait froid dehors, vous devez protéger votre corps.'
'Hommes et femmes ne sont pas égaux devant la douleur. Des études sur les animaux montrent que les mâles sont moins sensibles que les femelles. Des composants du corps jouant le rôle d'antidouleur seraient plus présents chez ces derniers.'

Table C.48 Keywords and duration for list 24

Keyword 1	Keyword 2	Keyword 3	Selected keyword	Extraneous keyword	Sec.
sept	perdure	infini	infini	illimité	11
interdit	véhicule	amoureux	interdit	défendu	11
beaucoup	partie	dehors	dehors	extérieur	11
égaux	sensibles	composants	sensibles	réceptifs	11

Appendix D

Résumé

Le phénomène Cocktail Party : aspects théoriques, comportementaux et neurophysiologiques

Il n'est pas toujours aisé de suivre une conversation dans un environnement bruyant. Il y a plus de soixante-cinq ans que Cherry a décrit ces situations de "Cocktail Party" et, encore aujourd'hui, nous n'en comprenons pas encore tous les mécanismes. Davantage comprendre ce phénomène est pourtant crucial à l'heure où se développent les moyens de communication à travers des connexions internet, qui peuvent parfois être instables et entraîner des difficultés de compréhension.

Les sons que nous entendons proviennent généralement de différentes sources acoustiques et arrivent à notre oreille sous la forme d'une mixture. Le premier défi auquel un auditeur fait face est de séparer cette mixture pour regrouper les différentes sources acoustiques. Une fois que les différents flux auditifs sont formés, l'auditeur doit, dans un second temps, concentrer son attention sur la source qu'il désire écouter. Un cas courant de l'effet cocktail party est la simple situation où, deux locuteurs, seulement sont en compétition ("parole dans la parole"). Bien qu'il s'agisse d'une version simplifiée de l'effet cocktail party, un travail considérable a été réalisé pour mieux comprendre comment les humains peuvent facilement se concentrer sur un locuteur ("cible") tout en ignorant l'autre ("masque").

Le premier chapitre de ce travail documente la littérature disponible sur la "parole dans la parole" à travers des approches comportementales et neurophysiologiques, comme le "speech tracking". Il apparaît que les méthodes comportementales emploient bien souvent des stimuli courts (syllabes, mots, phrases,

etc.) alors que les approches neurophysiologiques favorisent des stimuli longs de parfois plusieurs minutes. De surcroît, il est souvent supposé dans les études "speech tracking". que l'auditeur maintient une attention constante sur la voix cible, tandis qu'en situation réelle, l'attention des auditeurs peut rapidement varier d'une voix à une autre. Ainsi, ces différences dans les deux approches peuvent entraîner l'implication de différents mécanismes cognitifs, rendant les études comportementales difficilement conciliables avec un certain nombre d'études neurophysiologiques.

Afin de pallier ces limitations, un test d'intelligibilité – le Long-SWoRD – a été conçu. Aussi, le premier objectif de cette thèse est de documenter la performance de personnes normo-entendantes dans les situations où la difficulté de séparer perceptivement deux voix concurrentes varie de facile à difficile. Le second objectif est de faire le lien entre les études comportementales et neuropsychologiques d'une part, et d'autre part observer si les résultats obtenus en utilisant les deux approches sont cohérents les uns avec les autres.

Le Long-SWoRD test

Le développement et la validation d'une nouvelle tâche d'intelligibilité – le Long-SWoRD – sont introduits dans le deuxième chapitre. Le Long-SWoRD test est composé de brèves histoires que les participants doivent écouter attentivement pour ensuite retrouver, parmi un ensemble de mots-clés, ceux présents dans l'histoire cible.

Les histoires, pour la plupart des anecdotes, ont été extraites du livre audio " Le Charme discret de l'intestin" (Enders, Enders, & Liber, 2015), afin de maintenir au mieux l'attention des participants. Les mots-clés à retrouver à la fin de chaque histoire ont été sélectionnés afin d'éviter des effets de primauté, de récence ou encore de répétition. Par ailleurs, les mots-clés ont été analysés pour correspondre à une fréquence courante dans le langage. Chaque essai du Long-SWoRD test est composé de deux histoires, une histoire cible et une histoire masque, associées à leurs trois mots-clés respectifs.

La détection de trois mots-clés dans des histoires qui durent parfois jusqu'à 18 secondes semblait ardue pour les participants en raison de la charge cognitive de la mémoire de travail. Il a donc été décidé d'évaluer cette procédure expérimentale

par une étude en ligne. Les résultats ont permis de sélectionner 144 paires d'histoires, regroupées en 12 listes.

Pour finir, chaque paire d'histoires a été complétée avec trois mots-clés externes qui proviennent d'autres histoires. Comme toutes ces histoires proviennent du même livre et par extension d'un champ lexical similaire, il fallait s'assurer que les mots-clés externes n'étaient pas sémantiquement plus proches des mots-clés cibles ou des mots-clés masques. Afin que les mots-clés externes aient le moins d'influence possible sur le choix des participants, la similarité sémantique a été mesurée entre les mots-clés cibles, masques et étrangers.

En conclusion, le Long-SWoRD test permet, tout d'abord, de s'approcher de situations réalistes et, in fine, de bénéficier pour les participants de ressources cognitives, telles que des connaissances linguistiques, pour séparer deux locuteurs. Par ailleurs, ce nouveau protocole fournit aux expérimentateurs un moyen pour inférer à quel moment sur quelle voix les participants se concentrent.

Séparation de locuteurs avec le Long-SWoRD test

Dans le troisième chapitre, les performances des participants au Long-SWoRD test sont évaluées et documentées à travers deux études comportementales. Deux indices de ségrégation, largement étudiés et nécessaires pour démêler deux locuteurs, sont manipulés afin de contrôler la difficulté de la tâche. Ainsi, les histoires masques sont prononcées par le même locuteur que les histoires cibles, mais les paramètres vocaux (F0 et longueur du conduit vocal) ont été modifiés pour contrôler la similarité des deux voix. La difficulté de la tâche a également été modulée en alternant les modes de présentation des stimuli, à savoir des présentations dichotiques et diotiques.

Les résultats de la première expérience indiquent que les participants réalisent de meilleurs scores lorsque les voix sont présentées dichotiquement ou lorsque les voix sont suffisamment dissimilaires. Par ailleurs, lorsque les deux voix sont semblables, les sujets choisissent davantage les mots appartenant à l'histoire masque, ce qui refléterait des changements d'attention. Des analyses complémentaires indiquent que les participants peuvent également profiter du contexte sémantique dans les situations d'écoute compliquées.

La seconde expérience documente les performances des participants dans des situations d'écoute beaucoup plus difficiles que l'expérience 1. Dans ce contexte,

seuls les paramètres vocaux des locuteurs ont été manipulés et les stimuli sont présentés de manière diotique. De manière générale, les scores des participants diminuent proportionnellement avec la similarité des deux voix : au plus les voix sont proches, au plus les participants sélectionnent les mots-clés masques. L'analyse du contexte sémantique souligne, cette fois-ci, que lorsque les voix des locuteurs sont trop proches, l'apport de l'information sémantique est limité.

En conclusion, les résultats des deux expériences avec le Long-SWoRD test sont cohérents. De plus, tout comme dans la littérature, ils montrent que la distance entre les voix, les indices de spatialisation, ainsi que les informations sémantiques peuvent influencer les performances des participants.

Tracking neuronal avec le Long-SWoRD test

Traditionnellement, les études neurophysiologiques sont analysées avec des potentiels évoqués. Une des limites de cette approche est la difficulté d'analyser des stimuli d'une durée supérieure à quelques secondes. Plus récemment, les fonctions de réponse temporelle (TRFs) ont été développées pour contourner cette limitation. L'approche du "speech tracking" exploite ces TRFs afin de reconstruire le flux de parole sur lequel un auditeur se concentre. En général, ce champ de recherche montre qu'il existe une corrélation entre les oscillations corticales et des stimuli auditifs reconstruits à partir des TRFs.

Dans l'expérience présentée dans ce chapitre, l'activité cérébrale des participants est enregistrée avec un électroencéphalogramme (EEG) alors qu'ils passent le Long-SWoRD test dans différentes situations d'écoute. Tout comme dans les deux expériences présentées dans le chapitre précédent, la difficulté de la tâche est modulée par le degré de similarité entre les voix des deux locuteurs.

Les résultats comportementaux sont similaires aux deux expériences du chapitre précédent : lorsque les voix sont similaires, les participants font davantage d'erreurs. Les résultats neurophysiologiques montrent que l'histoire cible a une meilleure reconstruction que l'histoire masque. Par ailleurs, la différence entre ces deux représentations se réduit lorsque les voix sont similaires.

Les résultats comportementaux concernant le contexte sémantique sont également cohérents avec les deux expériences du chapitre précédent : les participants obtiennent de meilleurs scores au Long-SWoRD test lorsqu'ils peuvent s'aider du

contexte sémantique. Par ailleurs, la présence de marqueurs neuronaux confirme les résultats comportementaux.

Pour finir, la phase de séparation de la mixture acoustique en deux différents flux de parole est analysée au cours du temps : les représentations des histoires cibles et masques se construisent plus rapidement lorsque les voix sont facilement distinguables.

Optimisation comportementale du tracking neuronal

Pour parvenir à discriminer deux locuteurs, nous devons mobiliser de nombreux mécanismes perceptifs et cognitifs, ce qui peut parfois entraîner un basculement momentané de notre attention auditive sur les discussions alentour. Il est souvent supposé dans la littérature que l'attention des participants reste constamment sur la même voix. Le Long-SWoRD test permet d'inférer rétrospectivement à quel moment et quelle voix les participants écoutaient.

Dans ce cinquième chapitre est présentée une analyse combinée de ces informations attentionnelles et des signaux EEG du chapitre précédent. Sur base des réponses des participants au Long-SWoRD test, des stimuli reflétant l'écoute réelle des participants ont été modélisés pour ensuite être évalués avec une approche de "speech tracking".

Les résultats montrent que les informations concernant l'écoute réelle - par opposition à l'écoute supposée – des participants peuvent être utilisées avantageusement pour améliorer la précision de la reconstitution du stimulus. En particulier, dans les situations d'écoute difficiles où l'attention des participants est moins susceptible de rester entièrement concentrée sur l'interlocuteur cible. Dans les situations où les deux voix concurrentes sont clairement distinctes et facilement séparées sur le plan perceptif, l'hypothèse selon laquelle les auditeurs sont capables de rester concentrés sur la voix cible est raisonnable.

Discussion générale, conclusion et perspectives futures

Le premier objectif de cette thèse était de documenter la performance des personnes normo-entendantes avec le Long-SWoRD test. Tout d'abord, il est important

de souligner la cohérence des résultats comportementaux au travers des trois expériences présentées dans cette thèse. Ensuite, les résultats des participants reflètent, généralement, le même schéma que des performances mesurées dans de précédentes études de la littérature.

Le second objectif de cette thèse était de faire le lien entre les approches comportementales et neurophysiologiques. Les résultats comportementaux et neurophysiologiques de la troisième expérience présentée dans cette thèse soulignent la similarité des performances de ces deux approches. De plus, tout comme les résultats comportements, les résultats neurophysiologiques sont cohérents avec la littérature. Par ailleurs, la combinaison de ces deux approches a permis d'une part, l'optimisation des fonctions de réponses temporelles, et d'autre part, une contribution neurophysiologique à l'étude de la décomposition de la mixture en flux de parole au cours du temps.

Pour finir, l'influence de la mémoire de travail est soulignée dans ce dernier chapitre et de nouvelles perspectives sont envisagées à travers, notamment,, l'étude de "l'irrelevant sound effect".

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : HUET

DATE de SOUTENANCE : 17/09/2020

Prénoms : Moïra-Phoebé

TITRE : Voice mixology at a cocktail party: Combining behavioural and neural tracking for speech segregation

NATURE : Doctorat

Numéro d'ordre : 2020LYSEI070

Ecole doctorale : Mécanique, Energétique, Génie civil, Acoustique (MEGA)

Spécialité : Acoustique

RESUME : Il n'est pas toujours aisé de suivre une conversation dans un environnement bruyant. Pour parvenir à discriminer deux locuteurs, nous devons mobiliser de nombreux mécanismes perceptifs et cognitifs, ce qui peut parfois entraîner un basculement momentané de notre attention auditive sur les discussions alentour. Dans cette thèse, les processus qui sous-tendent la ségrégation de la parole sont explorés à travers des expériences comportementales et neurophysiologiques. Dans un premier temps, le développement d'une tâche d'intelligibilité – le Long-SWoRD test – est introduit. Ce nouveau protocole permet, tout d'abord, de s'approcher de situations réalistes et, in fine, de bénéficier pour les participants de ressources cognitives, telles que des connaissances linguistiques, pour séparer deux locuteurs. La similarité entre les locuteurs, et donc par extension la difficulté de la tâche, a été contrôlée en manipulant les paramètres des voix. Dans un deuxième temps, les performances des sujets avec cette nouvelle tâche est évaluée à travers trois études comportementales et neurophysiologiques (EEG). Les résultats comportementaux sont cohérents avec la littérature et montrent que la distance entre les voix, les indices de spatialisation, ainsi que les informations sémantiques influencent les performances des participants. Les résultats neurophysiologiques, analysés avec des fonctions de réponse temporelle (TRF), suggèrent que les représentations neuronales des deux locuteurs diffèrent selon la difficulté des conditions d'écoute. Par ailleurs, ces représentations se construisent plus rapidement lorsque les voix sont facilement distinguables. Il est souvent supposé dans la littérature que l'attention des participants reste constamment sur la même voix. Le protocole expérimental présenté dans ce travail permet également d'inférer rétrospectivement à quel moment et quelle voix les participants écoutaient. C'est pourquoi, dans un troisième temps, une analyse combinée de ces informations attentionnelles et des signaux EEG est présentée. Les résultats soulignent que les informations concernant le focus attentionnel peuvent être utilisées avantageusement pour améliorer la représentation neuronale du locuteur sur lequel est portée la concentration dans les situations où les voix sont similaires.

MOTS-CLÉS : Effet cocktail party ; Analyse scène auditive (ASA) ; ségrégation de locuteurs ; Attention auditive ; Tracking cortical ; Intelligibilité ; Voix ; Contexte sémantique ;

Laboratoires de recherche : Laboratoire Vibrations Acoustique (LVA)
Centre de Recherche en Neurosciences de Lyon (CRNL)

Directeur de thèse: Etienne Parizet

Président de jury : Daniel Pressnitzer

Composition du jury : Etienne Gaudrain
Carolyn McGettigan
Fanny Meunier