Language and its multi-level organization
How the brain puts order in the speech signal

by David Maximiliano Gómez Rojas

Submitted in partial fulfillment of the requirements for
Doctor Philosophiae degree in Cognitive Neuroscience

Supervisor: Jacques Mehler

Scuola Internazionale Superiore di Studi Avanzati (SISSA)
Trieste, Italy
December 2012

“The scientist must set in order. Science is built up with facts, as a house is with stones.
But a collection of facts is no more a science than a heap of stones is a house.”

— Henri Poincaré

This thesis in a nutshell

The complex process leading from the speech signal to its linguistic content involves the analysis of structures at different levels. The brain builds discourses from sequences of utterances, utterances from sequences of words, words from syllables, syllables from phonemes, and phonemes from acoustic features. This thesis is devoted to the study of the brain mechanisms that mediate this processing called *segmentation* in two different levels of organization: the building of words from syllables and the building of syllables from phonemes.

Experiments 1 and 2 explore the process of segmenting words from a continuous stream of syllables in adult participants. Specifically, they focus on the time course of word extraction using a behavioral method called click detection (Exp. 1) and a frequency analysis of electrophysiological data (Exp. 2). In particular, oscillatory neural activity suggests a dual-stage model of word segmentation, supporting models that claim that segmentation at this level proceeds by chunking syllables.

Experiments 3, 4, and 5 investigate syllabic perception by newborn infants as a means of assessing the structure of the *primordial syllable*: the perceptual unit of speech that is present before significant experience with language. Utilizing the linguistic concept of sonority and functional near-infrared spectroscopy, these experiments show: that activity in left temporal cortex distinguishes between sequences of phonemes yielding well- and ill-formed units (Exp. 3); that this difference is not due to low-level properties such as onset acoustics and signal envelope (Exp. 4); and that this difference vanishes if an alternative syllabic structure is possible (Exp. 5). These findings indicate that the construction of syllables from sequences of phonemes has specific constraints since the beginning of life.

The research herein described unveils fundamental aspects of the segmentation process in adulthood and infancy, contributing to our understanding of how the human brain transforms speech into language.

Acknowledgments

Éste es uno de esos momentos en los cuales las palabras y el lenguaje se hacen poco para expresar lo que siento. Y los idiomas comienzan a darse de codazos de modo de llegar primeros a la punta de mis dedos. English reclaims its supremacy because this paragraph introduces a research thesis, which is—after all—a formal work document. El Español sugiere, acertadamente, que es el vehículo más adecuado de agradecer a las personas más importantes de mi vida. E l'Italiano mi ricorda ad ogni istante tutta quella bella gente che è diventata la mia seconda famiglia, e la bellissima città di Trieste che mi ha accolto in questo periodo della mia vita. Me declaro incapaz de decidir. These three pull the strands of my mind, trying to capture and express my thoughts, feelings, and gratitude. Basta.

Being a scientific thesis the reason behind this document, let me start thinking about work. First and foremost I am grateful to my advisor Jacques Mehler, who some years ago believed in the potential of that young mathematician not only as a data analyzer but also as a scientist-to-be. Thanks also to Marina Nespor, who had an infinite patience to teach me about language, and specially about the melody of language (is there any motivation better than prosody for a musician to get interested into language?). I wish to thank to many former and current members of the Language, Cognition, and Development Lab of SISSA: Judit, Ricardo, Jean-Rémy, Amanda, Andrea, Alissa, Hanna, Alan, Ana, Yamil, and Erika among them. Some more, some less, all of you taught me many important things about science, about tenacity,

about friendship. Silvia: I simply could not write your name together with the rest. You have been an inspiring colleague and a great friend. Thanks for always being there, mae.

A special mention goes to Dr. Franco Macagno, whose invaluable help allowed me to complete a series of studies with newborn infants that compose a chapter of this thesis. Uncountable trips to Udine and back with Jacques, Silvia, Alissa, Ana, and Andrea were silent witnesses of our discussions about science, politics, and life. Sometimes these trips became the perfect excuse for napping. And they were an excellent opportunity for admiring the beauty of Venezia Giulia, an island between mountains and the Adriatic sea just like my homeland lies between the Andes and the Pacific.

Vorrei ringraziare pure Marijana Sjekloća, Francesca Gandolfo, Alessio Isaja ed Andrea Sciarrone. Senza di voi il caos della burocrazia mi avrebbe fermato dall'inizio di questo percorso. Grazie del vostro supporto, pazienza, cordialità e soprattutto della vostra amicizia. Especialmente tú, Scià: has sido un *coinquilino* excepcional. Gracias por dejarme entrar en tu casa (¡y por no sacarme a patadas en todos estos años!).

Gracias, papás, por su apoyo incondicional de todos estos años. Por dejarme partir libremente en búsqueda de un sueño, y siempre desearme lo mejor en los desafíos que decido afrontar. Gracias por su generosidad en todo sentido, y especialmente por darme la oportunidad de despedirme de la Carmen en persona. Juan Carlos y Camila, gracias por hacerme sentir siempre cerca a pesar de la lejanía. Gracias a todos ustedes por su cariño y apoyo. También a mis amigos del alma, Paula, Pati, Marcelo y Jairo, y a ti, Ángel.

Questa lettera di ringraziamento non può finire senza un paragrafo dedicato alla città di Trieste, e specialmente al particolare legame nato fra di noi. Perché mentre le neuroscienze hanno affascinato la mia mente, è stata la musica a nutrire il mio cuore. Il canto mi ha regalato una famiglia allargata e tanti amici, la cui lontananza ora mi spezza l'anima. Elia, Mira, Adriano, Walter: grazie per regalarmi dei momenti preziosi ed indimenticabili. Insieme ai vostri nomi dovrei includere un lungo elenco di amici che non oso scrivere per paura di dimenticarne qualcuno. Monica ed auricoralisti: grazie anche a voi, non solo per quanto riguarda la musica ma per essere diventati una seconda famiglia per me. Non riesco a decidere se mi mancheranno di più le nostre prove solite, o quelle alternative. Ragazzi cari, continuate a cantare tante belle cose! E vi prego, non smettete mai di far che le cose quotidiane diventino magiche.

There are many names that still come to my mind. My gratitude and best wishes go to Gloria, Raffaele, and Luka; a Paolo, Leonardo, Ivan, ed Alessandro e Fabio; to Antonino, Bailu, Sahar, Alireza, Shima, Ana Laura, Olga, Paola, Roma, Sarah, Gabriella, Federica, and many others.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Language acquisition is one of the most striking achievements of every healthy human infant. It involves the acquisition of a series of symbols at different levels, leading from the raw speech signal to phonemes, syllables, words, phrases, and utterances. Each one of these levels presents multiple regularities, constraints, and challenges for the infant brain. The origin of the linguistic knowledge helping infants to cope with this problem has prompted heated debates about the roles of inborn constraints and the powerful learning mechanisms available to infants, shaping one of the most important *nature* vs. *nurture* arguments in human cognition (e.g. see Crain & Pietroski, 2001). The present work assumes, following previous proposals from both inside (e.g. Fitch, 2011) and outside the language domain (Carey, 2009) that, in the end, any satisfactory explanation will consist of a mixture of these two extremes. We agree with Reali and Christiansen (2009) who propose, among others, that responses to the nature vs. nurture dichotomy for language will most probably arise from a multidisciplinary research program involving genetics, biology, neuroscience, cognitive science, linguistics, and other disciplines. This thesis work is partly based on linguistic theory, but its core—in terms of methodologies and argumentation—belongs to cognitive science and neuroscience.

In this introductory chapter, we first review some arguments for the necessity of inborn constraints for language: we start with the historical argument by Chomsky for the necessity of Universal Grammar for the acquisition of syntax, and then move on to other arguments coming from the scientific study of young infants acquiring their maternal language. We also present a section reviewing three important aspects of language development during the first year of life. Then we describe the main topic of this thesis, the problem of segmentation of continuous speech and the two-fold approach to it that we endorse. Lastly, we present short abstracts of the two main chapters of this thesis.

1.1 Inborn constraints in language acquisition

1.1.1 Chomsky's poverty of stimulus[1]

Noam Chomsky, in a series of influential publications, proposed that the task of acquiring a language is beyond what most infants can attain by means of their limited linguistic input. One of his paradigmatic examples asks how infants discover that language is hierarchically structured, rather than, say, linearly structured. To illustrate this, he considered the case of auxiliary fronting in English (e.g. Chomsky, 1971, 1975), which is a process that converts a declarative statement into a polar question:

Silvia is happy. → Is Silvia happy?

The guys are on holidays. → Are the guys on holidays?

The learning infant, after being exposed to sentences like these, might entertain at least two different hypotheses about how auxiliary fronting works. One option consists in building the polar question by moving to the beginning of the sentence the first instance of the verb

[1] This section is a brief introduction to the nativist view of language acquisition. For a comprehensive presentation, the reader is referred to Cowie (2010).

to be (linear hypothesis, LH). Alternatively, one could move to the beginning of the sentence the first instance of the verb *to be* that occurs after the subject of the sentence (hierarchical hypothesis, HH). Any scientist observing the infant's dilemma would agree that between these options, the former is more parsimonious, and it requires less complex processes and assumptions about the infant's mental capacities. However, the simple, non-hierarchical option turns out to be incorrect in the general case:

The person who is in the line is Andrea's brother.
[LH] → Is the person who in the line is Andrea's brother? ✗
[HH] → Is the person who is in the line Andrea's brother? ✓

Chomsky argued that correct exemplars of these complex transformations—the only ones carrying sufficient evidence to discard the non-hierarchical hypothesis—are scarce enough in infants' linguistic input so that infants would not be guaranteed to learn properly auxiliary fronting (hence the name, *poverty of stimulus*, of this argument). This implies, according to Chomsky, that facts like the hierarchical structure of language are not learned from experience, but rather are genetically encoded in humans in an abstract *Universal Grammar* (UG)[2].

As with many influential theories, Chomsky's proposal has been heatedly debated over the years. Even in our days, it sparks interesting research in fields such as computational modeling of language acquisition. For instance, Perfors, Tenenbaum, and Regier (2011) have suggested that complex syntactic structure can be inferred by the infant from simple exemplars, proposing a process based on Bayesian inference which might lead infants to learn auxiliary fronting in English interrogatives.

[2] The term *Universal Grammar* actually predates Chomsky's theories. Chomsky (1965) quoted James Beattie as describing UG in 1788 as a description of "those things, that all languages have in common, or that are necessary to every language". Chomsky, however, redefined UG as an object of study not only of Linguistics, but also of Biology. In Chomsky (2005), he refers to UG directly as the human "genetic endowment" for language.

Although Chomsky always acknowledged that universal commonalities of languages do not refer exclusively to syntax[3], he and other authors have proposed (Hauser, Chomsky, & Fitch, 2002) that recursion (a general cognitive process tightly linked to syntactic structure) is the only innate human specialization for language (also referred to as *Merge* in Chomsky, 1995, 2005).

1.1.2 The case of phonology

Stating recursion as the only innate human specialization for language—and thus, as the only component of UG—entails that all other properties and mechanisms of language are either consequence of general-purpose brain systems, or learned from experience. The linguistic and scientific communities received this conclusion with a mixed view, and critical reviews appeared soon. In 2005, one such critique by Pinker and Jackendoff stressed the importance and complexity of several other aspects of the language faculty, suggesting that the genetic endowment for language goes beyond recursion alone. An important part of Pinker and Jackendoff's critique can be illustrated by the observation that reducing the language faculty to recursion is like reducing grammar to syntax (see Berent, 2009, for a similar commentary). There is a broad consensus among linguists that grammar comprises morphology, phonology, and semantics as well.

Phonology studies the systematic organization of sounds in languages. It is related—but not limited to—allophonic variation, so for instance /p/ in *pin* is aspirated (pronounced [p^{h}ɪn]) whereas it is not in *spin* (pronounced [pɪn]), although /p/ is a unique phoneme in English. The nature of phonology is not recursive: phonemes cannot be embedded in other phonemes, and syllables cannot be embedded in other syllables. However, phonology seems

[3] For instance, consider the putative universal that "colour words of any language must subdivide the colour spectrum into continuous segments" (Chomsky, 1965, as cited by Fitch, 2011).

to be highly specific to language and some researchers claim that it presents regularities that apply universally across languages as well. One of these regularities will be studied in depth in Chapter 3.

1.1.3 Developmental cognitive science

A direct effect of Chomsky's work was the emergence of numbers of researchers from different fields, who focused their efforts on understanding the *initial state* with which infants come to the world. Although most of them were concerned with language, other topics such as object cognition and numerosity have been subject to thorough examinaton (see Carey, 2009, for a review on these topics).

Developmental cognitive science was then born as a field of study about the capacities and constraints that young infants display in their behavior. The advent of techniques based on infants' looking behavior and neural activity enhanced greatly the reach of developmental cognitive scientists.

Scholars in this field have demonstrated several interesting biases in infants' perception of language, showing that some aspects of the adult state are present since the very first weeks or days of life. For instance, human neonates adjust their sucking behavior so as to listen more to speech stimuli than to complex non-speech analogues (Vouloumanos & Werker, 2007); young infants discriminate some phonetic changes only when these are embedded in word-like units (e.g. they discriminate *tap* and *pat* but fail to discriminate *tsp* and *pst*, Bertoncini & Mehler, 1981); and they display larger mismatch responses for acoustic changes in speech if they cross a phonemic boundary than if they do not (Dehaene-Lambertz & Baillet, 1998), even though phonemic boundaries were originally defined on the basis of adult speech perception. Other findings attest that newborns infants have been able to learn intonational aspects of their maternal language while still in the womb (e.g. Mampe, Friederici, Christophe, & Wermke, 2009; May, Byers-Heinlein, Gervain, & Werker, 2011).

The study of language is a field that may benefit enormously from a developmental perspective. Human infants display an impressive facility for acquiring languages, in a seemingly effortless process. However, this facility vanishes in adulthood: learning a new language then becomes a difficult task, and results in terms of fluency or pronunciation are seldom comparable to native levels. The existence of this *critical* or *sensitive period* is probably due to maturational processes that modify the way our brains cope with linguistic stimuli (Newport, 2002), not limited to the auditory modality but including the visual modality as well (Mayberry & Eichen, 1991).

1.2 Language development in the first year of life

Infants in their first weeks and months of life deploy a wide range of mechanisms in order to acquire the language(s) of their surrounds. By the end of the first year of life most of them have already tuned to the phonemic repertoire of their native language, they utter their first words, and possess a sizable lexicon. This section presents a brief review of three important elements that are learned by normally-developing infants in their first year: the rhythm of their native language, its phonemic repertoire, and part of its lexicon. It is important to keep in mind that inter-individual variability in reaching developmental milestones may be big. This section thus describes abilities and knowledge acquired by typically-developing infants in their first year of life.

1.2.1 The native rhythm

Rhythm can be understood as the systematic patterning of sound in terms of timing, accent, and grouping (Patel, 2007). Languages also have a rhythm, and linguists have classified several of them according to their rhythmic properties. Dating back at least to the works by Abercrombie (1967) and Ladefoged (1975), languages of the world are thought to cluster in three main *rhythmic classes*:

Stress-timed languages. This class includes languages like English, Russian, and Dutch. Here, the main rhythmic unit is the *foot*, which in a first approximation can be considered as composed by a stressed syllable and all the unstressed syllables that follow.

Syllable-timed languages. This class includes languages like Spanish, Italian, and Cantonese Chinese. Here, the main rhythmic unit is the syllable.

Mora-timed languages. This class includes languages like Japanese, Ganda, and Tamil. Here, the main rhythmic unit is the *mora*, a subsyllabic unit.

Ramus, Nespor, and Mehler (1999; see also Nespor, Shukla, & Mehler, 2011) provided acoustic measures that allow to quantify this clustering. They showed that $\%V$, the proportion of time that is spent producing vowels during an utterance, and ΔC, the variability of the length of inter-vocalic intervals, captured the intuitive ideas of linguists about rhythmic classes. They also demonstrated that adult listeners can discriminate sentences from languages of different rhythmic classes just on the basis of their composition in terms of consonantal and vocalic intervals[4].

Infants are able to discriminate languages on the basis of their rhythmic properties since their first days of life. Nazzi, Bertoncini, and Mehler (1998) showed that newborn infants

[4] For an opposing view on the relevance of the rhythmic class for language discrimination, see White, Mattys, and Wiget (2012).

discriminate English from Japanese (belonging to different rhythmic classes), but not English from Dutch (belonging to the same rhythmic class). Moreover, these authors also showed that the cues infants use for discriminating these languages are not specific to any of them, because they discriminate sentences in English and Dutch from sentences in Spanish and Italian, but fail to do so when contrasting English and Italian versus Dutch and Spanish. Ramus (2002) showed that, as with adult listeners, the pattern of alternation between consonants and vowels seems to be enough to elicit discrimination in newborns.

By 4 to 5 months of age, infants display the first signs of processing preferentially their maternal language with respect to other languages in the same rhythmic class (e.g. Nazzi, Jusczyk, & Johnson, 2000; Peña, Pittaluga, & Mehler, 2010) and later, by 7.5 to 9 months, they prefer words that follow rhythmic patterns frequent in their maternal language (e.g. Jusczyk, Cutler, & Redanz, 1993; Jusczyk, Houston, & Newsome, 1999).

Although we have referred mostly to speech rhythm, intonation is also an aspect that infants grasp from a very young age. Indeed, newborn infants born to German- and French-speaking families have cry melodies resembling the intonational pattern of their maternal language (Mampe et al., 2009), suggesting that intonation was learned prenatally.

1.2.2 Phonemic repertoire

From their first month of life, infants react more to a given acoustic change when this change crosses a phonemic boundary (Eimas, Siqueland, Jusczyk, & Vigorito, 1971), suggesting that early phoneme perception is similar to adults in this aspect (see also the electroencephalographic evidence by Dehaene-Lambertz & Baillet, 1998 with 4-month-olds). Given that infants can potentially learn every language of the world, all together this means that they should discriminate every possible phonemic change, even those absent from their parents' language. Indeed, this turns out to be the case: Werker and Tees (1984) demonstrated that young infants discriminate these "non-native" phonemic contrasts, and that by the end of the first year they become *specialized listeners*, maintaining only the distinctions that are relevant in their linguistic environment (see also Kuhl et al., 2006). Maye, Werker, and Gerken (2002) proposed that this specialization is due to infants' sensitivity to the distributions of sounds in their environment: 6- to 8-month-old infants who are exposed to a unimodal distribution of exemplars drawn from the [da]-[ta] continuum lose temporarily the sensitivity to this change, whereas this does not happen with infants exposed to a bimodal distribution of exemplars. This explanation is consistent with the fact that infant-directed speech tends to exaggerate phonetic differences between phonemes (e.g. see Kuhl et al., 1997, for the case of vowels).

The process of perceptual specialization can be affected even by moderate exposure to a second language. Kuhl, Tsao, and Liu (2003) showed that although infants acquiring American English typically lose sensitivity to some contrasts of Mandarin Chinese, some hours of interaction with a native speaker of Mandarin brings back sensitivity to native-like levels. The authors also demonstrated that social interaction with the infant is a crucial element for recovery to happen, since sensitivity of infants exposed to recordings (audiovisual or just audio) of the same Mandarin speaker was not boosted.

1.2.3 The lexicon

Acquiring the lexicon of the maternal language progresses along two main lines during the first year. First, the infant faces the task of storing common word forms in memory, and then these word forms can become full-fledged words by adding meaning to them.

The first stage may start right after birth, subject to the constraints of an immature memory system (Benavides-Varela, Gómez, Macagno, et al., 2011). Behavioral evidence of infants recognizing highly frequent words such as their own name comes only by 4-5 months of age (Mandel, Jusczyk, & Pisoni, 1995). Just a month later, infants can make use of these word forms in order to parse continuous speech (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005).

Although the precise nature of these word form representations is not yet clear, in early stages it includes some speaker-specific information (Houston & Jusczyk, 2003) despite infants' precocious capacity for abstracting from speaker's identity in discrimination tasks (Dehaene-Lambertz & Peña, 2001). Intonational properties that are relevant to specific languages are incorporated as components of the word form representations by the second semester of life: for instance, infants born to English-speaking families consider word stress as relevant by 7 months (Curtin, Mintz, & Christiansen, 2005), but disregard certain aspects of intonation such as pitch and amplitude by 9 months (Singh, White, & Morgan, 2008).

By the end of the first year, the phonological content of word representations is quite detailed at least for consonants in onset position, as evidenced by studies exploring infants' reactions to mispronunciations (e.g. Swingley, 2005a). Moreover, one-year-olds display a bias for reliance on consonants—as opposed to vowels—in a word learning context (Hochmann, Benavides-Varela, Nespor, & Mehler, 2011).

Recent evidence gathered by Bergelson and Swingley (2012) suggests that infants enter the second stage—incorporating meaning to word forms—earlier than previously thought. By studying 6-9 month-old infants' eye gaze, these authors propose that infants understand the referential meaning of some words already at that age[5].

[5] Hochmann, Endress, and Mehler (2010) studied the opposite case, namely how do infants learn the non-referential nature of some words. Their work with 17-month-olds points to word frequency as a critical variable, with very frequent words being less likely than unfrequent words to be associated with an object in a word learning task.

1.3 Segmentation of continuous speech

Throughout this thesis, we will refer to segmentation in a broad sense to a process that forms units from a continuous stream. Segmentation processes are pervasive in language: a continuous speech stream can be parsed into phonemes; a continuous stream of phonemes can be parsed into syllables; a continuous stream of syllables can be parsed into words; and so on. In the literature, it is possible to find two related—though not identical—concepts of segmentation. The first comes from studies about rhythmic properties of languages, and consists on investigating what is the basic unit in speech perception. In our conception of segmentation, this corresponds to the organization of sequences of phonemes into syllables. The second concept, instead, assumes that speech is already organized in syllabic units and explores how these are grouped into words. Both notions of segmentation have been shown to depend to different extents on prosodic and statistical mechanisms, which we review in the following.

1.3.1 Prosodic segmentation and the formation of syllables

Prosody is the melody of language. As mentioned in Section 1.2.1, infants are sensitive to prosodic information even before birth. Because of this, early research on segmentation focused on how infants and adults relied on prosody, and more specifically on rhythm and intonation.

Their investigations on the basic perceptual unit of speech led Savin and Bever (1970) to propose that the syllable is perceptually preferred with respect to the phoneme, demonstrating that adult English listeners are faster to detect entire syllables (e.g. *bolf*) than to detect their initial phonemes (e.g. *b*). Later, Mehler, Dommergues, Frauenfelder, and Segui (1981) observed that adult French listeners detect the target sequence *pa* faster in words like *palace* than in words like *palmier*, whereas the opposite pattern of results occurs when detecting *pal*.

Given that *palace* and *palmier* are syllabified as [pa.las] and [pal.mje] respectively, they concluded that their results supported the primacy of the syllable as a perceptual unit. Yet, these results proved to be dependent on the rhythmic properties of French: Cutler, Mehler, Norris, and Segui (1986) showed that adult English listeners asked to perform the same task do not display a pattern of results consistent with the syllabic hypothesis. French is a syllable-timed and English a stress-timed language, suggesting that linguistic rhythm affects speech perception. Extra support for this hypothesis came from analog findings in Japanese, a mora-timed language. That is, in Japanese the mora turns out to be the basic perceptual unit (Otake, Hatano, Cutler, & Mehler, 1993). All together, these findings indicate that adult listeners deploy segmentation processes based on the rhythmic characteristics of their native language.

But, as reviewed in Section 1.2.1, the acquisition of the rhythmic characteristics of the maternal language is carried out during the first months of life. The basic perceptual unit before significant exposure to language must be therefore independent of linguistic exposure, and in particular independent of the rhythmic class of the maternal language.

There is a number of studies that suggest strongly that newborn infants do not perceive speech discretized in terms of phonemes. Bijeljac-Babic, Bertoncini, and Mehler (1993) showed that neonates discriminate lists of bisyllabic and trisyllabic words composed of 4 and 6 phonemes respectively (e.g. *rifo*, *mazopu*), but fail to discriminate lists of bisyllabic words composed of 4 and 6 phonemes (e.g. *rifu*, *alprim*). Working with infants younger than 2 months, Bertoncini and Mehler (1981) demonstrated that syllable-like phoneme sequences are processed better than arbitrary phoneme sequences: Whereas they readily discriminate *pat* and *tap*, infants do not show any sign of discriminating *pst* and *tsp*. Further support for the role of the syllable came from a control experiment of the same work, in which infants succeeded in discriminating *utspu* and *upstu*, demonstrating that the lack of discrimination for *tsp* and *pst* is not due to poor phonetical discrimination or an inability to access phonemes in middle positions. Additionally, infants aged 4 months or less seem to fail to extract a phoneme-based regularity from a set of syllables (e.g. the common onset consonant in the sequence *bi*, *bo*, *ber*, *ba*, Jusczyk & Derrah, 1987), but they do extract similar syllable-based regularities (Eimas, 1999).

Later, Bertoncini, Floccia, Nazzi, and Mehler (1995) addressed directly the question whether neonates' basic speech perception unit is the syllable or the mora. With a design similar to Bijeljac-Babic et al. (1993), they presented neonates born to French-speaking families with bisyllabic and trisyllabic words (e.g. *iga*, *hekiga*) or with bisyllabic words composed by 2 and 3 morae (e.g. *iga*, *iNga*). Again, infants discriminated the sets of words that differed in number of syllables, and failed to discriminate the sets of words that had the same number of syllables (despite of having a different number of moras).

Prosodic cues have also been proposed for segmentation of words. Cutler and Norris (1988) proposed a *metrical segmentation strategy* specific to English. Since about 90% of English content words are stress-initial in conversational speech (Cutler & Carter, 1987), these authors proposed that an English learner segments words by assuming that stressed syllables are initial.

1.3.2 Statistical segmentation and the formation of words

Statistical segmentation refers to the mechanisms used to extract words from continuous streams of syllables based only on statistical cues such as token frequency and transitional probabilities[6].

[6] The *transitional probability* between units x and y refers to the probability that y follows x in a given stream of speech. Denote by $\#x$ the number of occurrences of the unit x in the stream, and by $\#xy$ the number of occurrences of the bigram xy. Then,

$$TP(x \to y) = \frac{\#xy}{\#x}.$$

Natural speech never presents statistical cues in isolation, because intonation is pervasive in natural speech. Nonetheless, Goodsitt, Morgan, and Kuhl (1993) reported that infants aged 7 to 8 months tend to follow statistical cues interior to intonational contours even when pitch is kept as flat as possible. Recent technological developments permitted the synthesis of good quality speech streams controlling to a far greater extent not only the pitch contour but also the duration of each phoneme (e.g. the MBROLA synthesizer, Dutoit, Pagel, Pierret, Bataille, & Van Der Vreken, 1996), providing a much finer control of intonation contours.

In a series of widely cited studies, Saffran and colleagues (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) reported that both infants and adults are able to extract words from monotone, meaningless, artificial speech streams without pauses or other intonational properties that might have signaled word boundaries. Based on their results, they argued that participants succeeded in parsing such streams because of their sensitivity to the transitional probabilities (TPs) between syllables[7]. Later studies have shown that these statistical computations are both domain (e.g. Saffran, Johnson, Aslin, & Newport, 1999) and modality (e.g. Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002) general.

Statistical computations are typically considered to be calculated at the level of the syllable. At least for the simplest syllable type (consonant+vowel, e.g. *ba*, *di*), this holds true even for adult listeners of languages from different rhythmic classes (e.g. see Saffran, Newport, & Aslin, 1996 for English listeners; Tyler & Cutler, 2009 for English, French, and Dutch; and Toro, Sebastián-Gallés, & Mattys, 2009 for English, French, and Spanish). Some groups have evaluated whether other, smaller, units are useful as well. The studies on vowels and consonants carried out by Bonatti, Peña, Nespor, and Mehler (2005, 2007) have shown that

[7] The work by Aslin, Saffran, and Newport (1998) intended to be a demonstration that infants are actually computing TPs instead of other measures such as simple bigram frequency. However, one of their main methodological assumptions turned out to be flawed (French & Perruchet, 2009), leaving the issue open again. In practice, when designing and analyzing our experiments we will consider TPs as the relevant measures, keeping in mind that several other possible measures may do the job equally well.

TPs between consonants are readily used for segmentation purposes, whereas TPs between vowels are not.

The ability to track statistical regularities in perceptual input is thus a general mechanism available for parsing continuous speech. Teinonen, Fellman, Näätänen, Alku, and Huotilainen (2009) suggest that this ability is present from the first days of life[8].

Recent studies have also examined the limits of segmentation based on statistical cues. For instance, most studies have used artificial languages in which all words have the same number of syllables. This extra regularity has proved to be relevant for segmentation performance in infancy, for infants do not segment correctly a stream composed of bi- and trisyllables after short exposure (Johnson & Tyler, 2009, but see Tyler & Cutler, 2009 for a similar design in adult research that gives positive results). Moreover, statistical segmentation of streams with uniform word length is affected by priming with words of the same or a different length (Lew-Williams & Saffran, 2012).

Analysis of corpus data has not been conclusive with respect to the usefulness of statistical mechanisms[9]. Gambell and Yang (2006) analyzed child-directed English speech and found little support for the success of statistical cues alone. Performance was greatly enhanced when assuming Cutler and Norris's (1988) English-specific metrical segmentation strategy (see Section 1.3.1) as a previous step to statistical computations. However, similar analyses conducted by Swingley (2005b) suggest the opposite: statistics may be enough not only for correct segmentation, but also for learning the aforementioned metrical strategy.

[8] It is relevant to point out that Teinonen and colleagues worked with a speech stream composed by syllables separated by pauses. It is not clear whether this *syllable grouping* task is entirely equivalent to segmentation, which involves also the extraction of a fixed set of phonemes from a constantly changing auditory stream (because of phenomena such as coarticulation).

[9] Pelucchi, Hay, and Saffran (2009) reported that infants succeed in segmenting a pair of Italian words from naturalistic stimuli. However, their material consisted on a very simplified corpus prepared with the aim to provide statistics about specific words and part-words. Their experiment cannot, thus, be considered a full-fledged test of the usefulness of statistics in a natural language.

A final piece of evidence supporting the relevance of statistical segmentation for lexical acquisition is the recent study by Singh, Reznick, and Xuehua (2012), who observed that infants' performance at a statistical segmentation task at 7 months of age correlates with measures of productive vocabulary size at 24 months.

1.3.3 Segmentation based on multiple cues

In recent years, both the importance of statistical cues and their insufficiency as the only basis for segmentation have been widely discussed and acknowledged. Many groups have thus explored how different segmentation mechanisms interact with each other, especially with statistical computations.

For instance, the presence of isolated or highly frequent words also provides strong cues that affect segmentation from flat streams or naturalistic stimuli (Bortfeld et al., 2005; Brent & Siskind, 2001; Lew-Williams, Pelucchi, & Saffran, 2011).

We already mentioned Goodsitt et al.'s (1993) work, who pointed out that infants recur to statistics to parse the interior portions of utterances defined by prosody. In a similar line, Shukla, Nespor, and Mehler (2007) tested Italian adult listeners' capacity to extract words based on statistical cues when these cues collide with prosody: they found that phrasal prosody acts as a gate, allowing only some statistical cues to be extracted. Other researchers have explored how statistics interact with acoustic markers of word stress, such as pitch and duration (e.g. Ordin & Nespor, under review; Toro et al., 2009; Tyler & Cutler, 2009). Bion, Benavides-Varela, and Nespor (2011) demonstrated that pitch alternations are sufficient to elicit segmentation of a stream with no statistical cues at 7 months of age.

1.4 Outline of this thesis

This work comprises two main chapters dealing with the process of segmentation in two different levels of processing, coinciding with the two aspects of segmentation described in this introductory chapter. Chapter 2 explores segmentation as a process for word extraction, and Chapter 3 investigates early constraints on the syllable, as a proxy to the early processes extracting syllables from sequences of phonemes. Then, Chapter 4 presents a general discussion of this work. A brief abstract of each one of the two main chapters follows below.

1.4.1 Segmenting words from syllabic streams

Studies on word segmentation from continuous speech have typically addressed it as an all-or-none process: Participants have to segment a speech stream presented during an exposure phase, and perform a behavioral test afterwards. Being statistical segmentation an inherently dynamic process, it is striking how few studies have explored its time course along an experimental session. This apparently simple change leads to several methodological difficulties both for behavioral and psychophysiological studies. We review the literature on monitoring tasks and electrophysiological studies of segmentation, and propose methods to overcome these issues. We then present two experiments with adult participants, one recording behavioral measures and one recording electroencephalography. Both of them reveal successful segmentation of the speech stream by the third minute of exposure, and moreover they give insights on the processes underlying word segmentation.

1.4.2 Sequences of phonemes and the primordial syllable

Adult listeners parse speech based on units that are dependent on their native language, and more specifically on the rhythm of their native language. However, empirical evidence indicates that before infants acquire the rhythmic properties of their maternal language they perceive speech as organized into syllables. We ask whether newborn infants' notion of syllable is constrained by similar laws to adult syllabic perception. Specifically, we introduce the notion of sonority and the Sonority Sequencing Principle, which linguists have proposed as a universal restriction on languages' syllabic structure. We present three experiments with neonates, whose brain responses were assessed by means of functional near-infrared spectroscopy. Results exhibit a consistent hemodynamic pattern, with channels located over left temporal cortex responding differently to well- and ill-formed syllable types.

Chapter 2

Online evolution of word segmentation

Segmentation of continuous speech is a fundamental process for learning the lexicon of any natural spoken language, where silent pauses delimiting words are the exception rather than the rule. Yet, the human infant manages to extract common words quickly during the first months of life. Bergelson and Swingley (2012) claim that 6- to 9-month-old infants have already not only learned several words that are frequent in their environment but have also associated them to their referents, implying that segmentation starts strikingly early in life. Other studies establishing segmentation in naturalistic settings by this age have been conducted by Jusczyk and Aslin (1995); Tincoff and Jusczyk (1999).

As soon as their lexicon has a few words, infants can use them to infer boundaries of novel words (Bortfeld et al., 2005), initiating them in a bootstrapping process with exponential growth. However, this picture does not make clear how the seeds of this process are planted. With high probability, intonation delimiting phrase boundaries, words uttered in isolation, rhythmic properties of the maternal language, and statistical information concur to this aim (e.g. Cutler & Norris, 1988; Goodsitt et al., 1993; Lew-Williams et al., 2011).

2.1 Approaching the learning process

The search for efficient methods to understand the unfolding of mental processes like segmentation is crucial for the progress of cognitive science. Offline methods provide a measure of performance after the process has finished. Designs consisting on a familiarization phase and a subsequent test phase are representative of this category. An archetypical example in developmental cognitive research is the assessment of learning by means of the head-turn preference procedure (e.g. Gomez & Gerken, 1999). Alternatively, online methods provide a series of performance measures, each corresponding to a specific moment in time during the experimental task. Typically requiring more complex designs than their offline counterparts, online methods are scarce in developmental research. Examples in adult and animal research are serial reaction time tasks (e.g. Hunt & Aslin, 2001; Papachristos & Gallistel, 2006), which need large numbers of trials to give informative results.

Offline methods may be particularly successful for qualitative questions: whether a given capacity or process occurs or not under some conditions. Online methods are best suited for quantitative questions about the temporal unfolding of a mental or neural process.

In this chapter, we approach the process of word segmentation from continuous speech from a behavioral and neural point of view, aiming to obtain online indexes of segmentation. In Experiment 1, we borrow a behavioral method used in psycholinguistics to investigate online sentence processing, namely response latencies to detect clicks. Experiment 2 will explore the neural correlates of segmentation by looking at the time evolution of spectral power during the exposure to continuous speech. Special emphasis will be given to how the time evolution of behavioral and neural measures shed light on the nature of the segmentation process.

2.2 Behavioral approaches

Most behavioral studies approaching word segmentation have assessed word extraction using offline methods. The standard use of two-alternative forced choice tests with adults and the head-turning procedure with infants has prevented researchers from investigating the time course of the segmentation process. Nevertheless, understanding the time evolution of this process might be as interesting for research as its end product, that is to say the identification of segmented words when provided the aforementioned tests. A detailed study of the time course of segmentation could reveal constraints relevant for any candidate model (specially computational models) of word segmentation.

Furthermore, the mainstream methods have left some unresolved issues regarding the amount of exposure that is required before segmentation arises with different materials and age groups. Adults have generally received 10 minutes or longer times of exposure before they are tested, whereas infants have been exposed to the speech stream for just two minutes before the test. Some studies like the one conducted by Peña, Bonatti, Nespor, and Mehler (2002) have used short exposure periods with adult participants, but their material included subliminal pauses between words since their interest was to assess rule learning rather than segmentation. Endress and Bonatti (2007) proposed that rule learning is usually carried out on the basis of sparse data, whereas statistical computations take more time to produce above-chance performance in offline tests. Do adult participants actually require extra exposure for statistical computations, or is it just a consequence of the selected testing methods?

2.2.1 Monitoring methods and the perceptual displacement hypothesis

Some decades ago, psycholinguists devised the click location paradigm in an attempt to investigate sentence processing. In this method, participants are asked to listen to a sentence and then report the location of a click superposed to it. Fodor, Bever and Garrett used this method extensively (e.g. Fodor & Bever, 1965; Fodor, Bever, & Garrett, 1974; Garrett, Bever, & Fodor, 1966) to study syntactic processing. Concurrently, other methods such as phoneme detection, word detection, click detection and tone detection were also employed. A common underlying assumption was that reaction times (from now on, RTs) to detect the corresponding targets reflect the complexity of sentence processing in real time. Some of the early studies using these methods showed, for instance, that RTs to phonemes are longer in structurally complex sentences as opposed to structurally simple ones (Foss & Lynch, 1969) and that RTs to clicks located in major syntactic boundaries are shorter than RTs to clicks not in a boundary (Holmes & Forster, 1970).

Morgan (1994) adapted click detection to study the role of distributional and rhythmic cues for word segmentation in 8-month-old infants (see also Morgan & Saffran, 1995). Babies were trained to turn their heads to a location above a loudspeaker (where a puppet would appear) each time a buzz occurred during a stream of words. In a testing session, babies oriented faster to the puppet location after buzzes external to cohesive units defined by the aforementioned cues than after buzzes internal to these units.

2.2.2 Shortcomings

Findings obtained from different monitoring tasks were not always consistent (for a review of the early results of detection methods see Cutler & Norris, 1979). For example, RTs to phonemes in stressed syllables are shorter than to those in unstressed syllables, whereas the reverse effect holds for RTs to clicks (Bond, 1972; Cutler & Foss, 1977).

In a groundbreaking work, Cutler, Kearns, Norris, and Scott (1993) found that monolingual English and French listeners showed the same pattern of responses when detecting clicks within English sentences (only intelligible to the English listeners). Based on these results, they argued that click detection is primarily sensitive to acoustic properties of the linguistic material rather than to more abstract properties such as syntax or semantics.

Given the strong dependence on prosody of results given by monitoring methods, only a few investigations having well-controlled material could be expected to reflect real sentence comprehension. Two examples of these are the works conducted by Frauenfelder, Segui, and Mehler (1980) and Cohen and Mehler (1996). In both studies, the key feature driving click RTs was the reversibility of object and subject functions in relative clauses: compare, for instance, the sentences "Le garçon qui vit la fille" ["The boy who saw the girl"] and "Le garçon que vit la fille" ["The boy whom the girl saw"]. This is an example taken from Cohen and Mehler (1996), where by changing only one phoneme (/i/ to /ə/, and viceversa) they reversed object and subject. Frauenfelder et al. (1980) showed that reversibility of a relative clause induces different RTs to phonemes located after the clause boundary for subject and object relatives. Cohen and Mehler (1996), using click detection, replicated and extended these results: the RT difference is specific to transposed instead of normal object relatives, and to relatives instead of active or passive sentences.

Still, the confound with prosody reduced importantly the range of applicability of monitoring methods with naturalistic stimuli.

2.2.3 An opportunity for online analysis

Despite their important shortcomings, monitoring methodologies may still prove useful in contexts where prosody can be tightly controlled. This is the case with synthesized speech, where intensity, duration, and pitch of each segment can be manipulated with precision. A monitoring method may thus be used to contrast response latencies to elements occurring in unit boundaries or within units. From all the monitoring techniques, we selected click detection: clicks represent a simple way of inserting an auditory signal in an arbitrary location of a continuous speech stream, just like the buzz noises used by Morgan (1994). Experiment 1 will evaluate the informativeness of click detection for obtaining a segmentation index.

2.3 Psychophysiological approaches

More recently, neuroscientists studying the word segmentation process have devised a variety of designs to approach it with neuroimaging.

In general, these studies coincide in that the neural mechanisms underlying statistical computations and segmentation yield evidence of learning several minutes before performance in offline tests can be assessed.

2.3.1 Electro- and magneto-encephalography (EEG/MEG)

The first study about segmentation from artificial, synthesized, continuous speech was conducted by Sanders, Newport, and Neville (2002). They measured ERPs to the presentation of nonce words and part-words (that is, sequences of syllables that occur during an artificial speech stream but that do not correspond to the words themselves) before and after exposure to an artificial speech stream. They reported an enhanced N400 component[1] for words with respect to part-words after exposure, and an enhanced N100 component only for the better learners according to a behavioral test.

Abla, Katahira, and Okanoya (2008) and Abla and Okanoya (2009) measured ERPs while participants listened to a stream of pure tones or observed a stream of geometrical figures, respectively. In both studies, they found an enhanced N400 component at middle frontal

[1] Associated to semantic priming, e.g. see Bentin, McCarthy, and Wood (1985).

and central scalp locations for "word" instead of "part-word" onsets[2] in the participants who performed better in a behavioral offline test. This ERP pattern was present during the first minutes of continuous exposure to the statistical streams only.

Other researchers, like De Diego Balaguer, Toro, Rodríguez-Fornells, and Bachoud-Lévi (2007) and Buiatti, Peña, and Dehaene-Lambertz (2009) have mainly approached a problem similar to segmentation, which is how adults discover non-adjacent dependencies in continuous speech. Both groups used an AXC paradigm, where a continuous speech stream is composed by trisyllabic words in which the first syllable predicts with certainty the third one. In an analysis that regarded word extraction rather than learning the non-adjacent dependency, De Diego Balaguer et al. (2007) also obtained a significant N400 effect evoked by word onsets that arises between the first and second minute of exposure in central electrodes. Rodríguez-Fornells, Cunillera, Mestres-Missé, and de Diego Balaguer (2009) later interpreted this N400 effect to the "progressive enhancement of a proto-lexical memory trace for the repeatedly encountered new word" (p. 3715).

Buiatti et al. (2009) measured EEG oscillatory power while participants listened to statistically or randomly structured speech streams. They observed a decrease of oscillatory power at the frequency corresponding to single syllables for structured speech streams compared to random ones. In addition, when they exposed participants to speech streams composed of trisyllabic sequences separated by subliminal pauses, they found greater oscillatory power at the frequency corresponding to trisyllabic sequences in streams composed by words than in random ones. These effects could be appreciated at the end of the third minute of exposure.

Teinonen et al. (2009) and Teinonen and Huotilainen (2012, MEG) have also investigated the segmentation process, although they did not try to establish a time course for the pro-

[2] The quotes are used to emphasize that in these studies the material was not linguistic, so that it is not possible to speak properly about words and part-words. In this case, what is meant is their respective analogs, namely cohesive units with high or low transitional probabilities.

cess. Particularly noteworthy is their 2009 work, which suggests that newborn infants can use transitional probabilities to group syllables[3] (see also Kudo, Nonaka, Mizuno, Mizuno, & Okanoya, 2011, for a study of segmentation of tone sequences with newborns).

2.3.2 Functional magnetic-resonance imaging (fMRI)

Although not directly relevant to the time evolution of segmentation, works using fMRI have investigated the neural bases of the segmentation process. By presenting blocks of artificial speech streams composed of either words or randomly concatenated syllables to adult participants, McNealy, Mazziotta, and Dapretto (2006) found that word structure elicited higher activations than random syllables bilaterally in temporal cortex. Additionally, these authors observed that activity in left superior temporal gyrus (STG) for the structured stream (compared to the random one) correlated with participants' behavioral accuracy. McNealy, Mazziotta, and Dapretto (2009) extended this investigation to 10-year-old children, replicating the activation in STG (with a reduced magnitude with respect to adults). A left STG activation for structured streams was also replicated by Cunillera et al. (2009).

Turk-Browne, Scholl, Chun, and Johnson (2009) used visual stimulation, finding that lateral occipital and left ventral occipito-temporal cortices were more activated for statistically structured as compared to randomly structured blocks. They also found several other brain regions that displayed activation preferentially for statistically structured blocks after two or three minutes of exposure. However, left STG was not activated for structured visual streams, suggesting that the neural bases of segmentation are modality-dependent.

[3] It is a subtle discussion whether the studies by these authors can be considered as real segmentation, since they worked with streams of syllables separated by pauses. Can segmentation be reduced to grouping of adjacent syllables? Although this simplification overlooks phenomena like coarticulation, it seems legitimate in the light of evidence showing that the syllable is the perceptual unit of speech (see Section 1.3.1).

2.4 Experiment 1

Time course of segmentation

Our first experiment used click detection as a behavioral measure to approach the time course of segmentation of continuous speech.

2.4.1 Participants

Twenty-eight adults participated in this experiment (11 men and 13 women, aged 23 ± 3 years). They were paid for their participation, reported no auditory or language-related problems and were naive with respect to the aims of this study. All participants were native speakers of Italian, recruited from the city of Trieste.

2.4.2 Stimuli

An artificial speech stream was generated using the MBROLA speech synthesizer (Dutoit et al., 1996) and the Italian female diphone database it4[4], with a sampling frequency of 16 KHz. The initial and final 5 seconds were ramped to avoid extra cues to segmentation. The stream contained the pseudowords pabuda, gifoto, venola, minaro. Each syllable lasted 240 ms, with no pauses between consecutive pseudowords. Word order was randomized, avoiding adjacent word repetitions. Using the software Praat (Boersma, 2001), a set of clicks was inserted into the speech stream by modifying the audio waveform, clipping five consecutive samples of it. As it can be seen in Figure 2.1, each click could occur either between two words (e.g. pabuda!gifoto, where the exclamation mark indicates the location of the click) or within a word, immediately after the first syllable (e.g. pa!buda). The stream was 4 minutes

[4] This database was created by the Istituto Trentino di Cultura, ITC-irst (downloaded from `http://tcts.fpms.ac.be/synthesis/`).

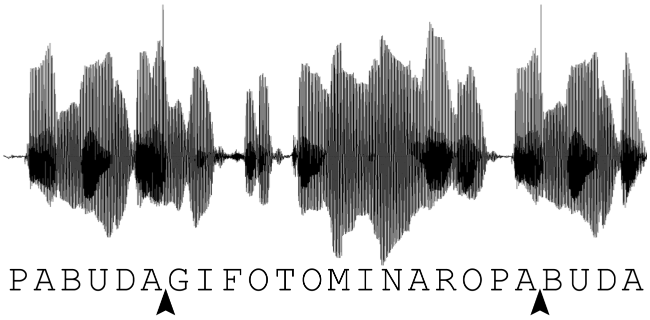

Figure 2.1: Example waveform of one of the artificial speech streams. The arrows show a click located between words (*pabuda*!*gifoto*) and a click located within a word (*pa*!*buda*).

long and contained 64 randomly spaced clicks, with an average interval of approximately 3.81 seconds (*SD* = 1.43, range 1.68–11.28) between consecutive clicks. A second speech stream was also generated, containing the words dagifo, nolami, narove, topabu and different click positions, in order to control for acoustic properties of the particular items and possible timing cues due to the chosen click positions (average interval between clicks was 3.90 for this stream, *SD* = 1.45, range 1.68–11.28). The words of the second stream were obtained by selecting "part-words" of the first, i.e. sequences of syllables that occur in the stream when concatenating syllables corresponding to different words. Half of the participants were exposed to each stream.

2.4.3 Procedure

Participants were tested individually in a silent room, and listened to the material through headphones. They were instructed both to attend the speech stream (a "discourse in an alien language") trying to extract the words it contained, and to press a key as fast as possible each time they heard a click. One RT for each click was obtained. Measurements were obtained using PsyScope X Build 45 (`http://psy.ck.sissa.it/`).

2.4.4 Data processing and analysis

RTs longer than 1,000 ms or shorter than 100 ms were excluded from any analysis[5]. RT data was then logarithmically rescaled for statistical analysis (descriptive statistics, ANOVAs and paired t-tests), rejecting for each participant all the rescaled RTs deviating more than 3 standard deviations from the average (leading to the rejection of less than 2% of the data). All graphics show the values converted back to normal scale, and the depicted 95% confidence intervals were adjusted, for a better visualization of the within-subject comparisons when required, using the method proposed by Cousineau (2005) corrected by Morey (2008). Significance of post-hoc analyses was corrected in accordance with the Holm-Bonferroni method (Holm, 1979).

All analyses were carried out using MatLab 2008b (Mathworks, Inc.) and SPSS 11 (SPSS, Inc.).

2.4.5 Results

Different RTs for both click locations

Figure 2.2a depicts participants' average RTs for both speech streams. A repeated measures ANOVA with between-subjects factor Stream (1 or 2) and within-subjects factor Click Location (Between words or Within words) reveals a significant effect of Click Location, $F(1,26) = 37.59$, $MSE = 0.04$, $p < .001$, and no significant effect of either Stream, $F(1,26) < 1$, $MSE < 0.01$, or interaction, $F(1,26) < 1$, $MSE < 0.01$. Thus from now on we work with the pooled data of all participants.

[5] This first filter was applied to RTs in order to detect anticipations and non-responded clicks. Values detected by this filter were typically one order of magnitude larger than median RTs (e.g. 3 s versus 300 ms). These strong differences advised against the use of a unique filtering step based on standard deviations.

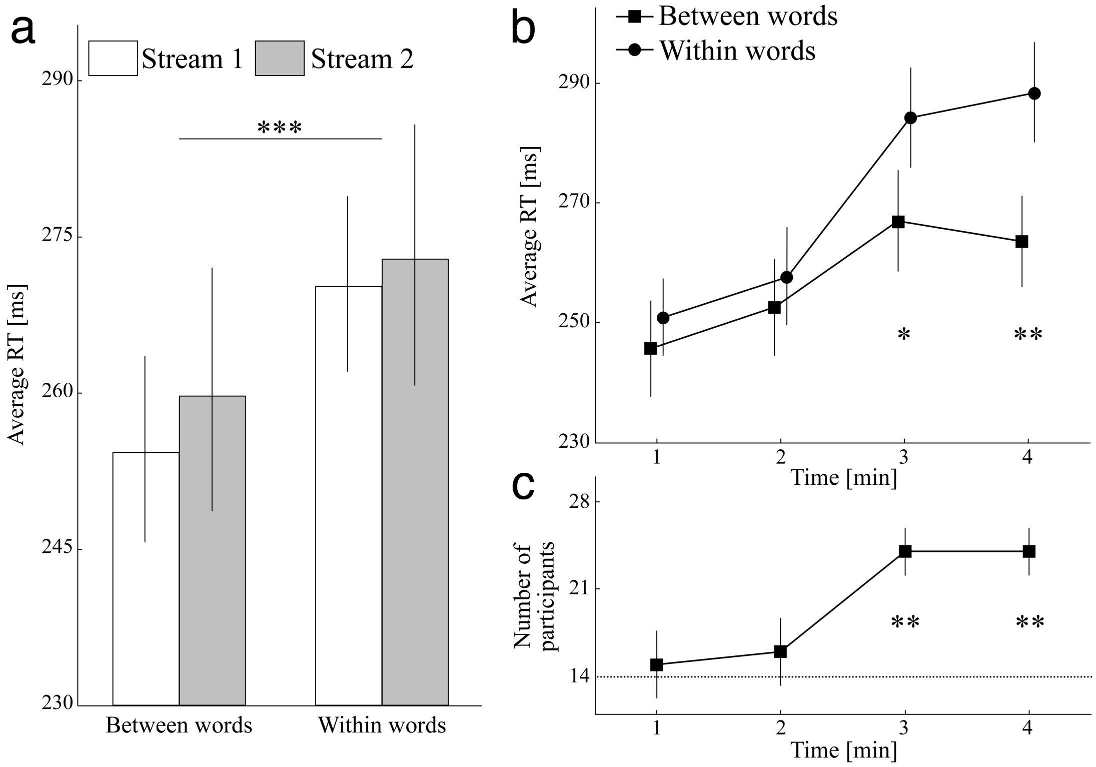

Figure 2.2: (a) Average RTs to clicks, segregated by Stream (1 or 2) and Click Location (between or within words). (b) Average RTs for both click locations, pooling separately RTs for each minute. (c) Number of participants for whom the average RT to clicks within words is longer than to clicks between words. The dotted line represents the chance level of 50% (14 out of 28). Vertical lines represent 95% confidence intervals, adjusted for within-subject comparisons when necessary (see Section 2.4.4). ** $p < .01$, *** $p < .001$.

Time evolution of RTs

To further analyze the time-course of this RT difference, in Figure 2.2b we present average RTs computed minute by minute. A two-way repeated measures ANOVA with factors Click Location (Between words or Within Words) and Minute (1, 2, 3 or 4) shows again a main effect of Click Location, $F(1,27) = 26.57$, $MSE = 0.13$, $p < .001$, with RTs to clicks located within words being longer than RTs to clicks located between words. We also observe a main effect of Minute, $F(3,81) = 16.99$, $MSE = 0.16$, $p < .001$ and a significant interaction,

$F(3,81) = 4.04$, $MSE = 0.02$, $p < .01$. Put together, these effects reveal that RTs to clicks located between words do not differ significantly across time (a one-way repeated measures ANOVA considering only the RTs for clicks located between words shows no significant effect of Minute, $F(3,81) = 1.55$, $MSE = 0.04$), whereas RTs to clicks located within words are longer in the second half of the exposure ($t(27) = 7.92$, $p < .001$, paired t-test for average RTs in minute 1 and 2 versus 3 and 4). Regarding the RT differences between both click locations, a post hoc analysis shows these are non-significant during the first half of the exposure and significant in the third and fourth minutes (paired t-tests for minute 1: $t(27) < 1$, *n.s.*; minute 2: $t(27) = 1.43$, *n.s.*; minute 3: $t(27) = 3.88$, $p < .01$; minute 4: $t(27) = 5.09$, $p < .001$). Since no other information apart from the transitional probabilities between syllables could cue the different click locations across both streams, this result suggests that the difference in RTs is due to the discovery and perception of the statistical words present in each stream through the segmentation process.

A similar pattern appears if we consider the number of participants for which the average RT associated to clicks located within words is longer than that associated to clicks located between words. As can be seen in Figure 2.2c, in minutes 3 and 4, 24 out of 28 participants were faster responding to clicks located between words than to clicks located within words ($p < .001$, binomial tests), in contrast with 15 out of 28 for minute 1 and 16 out of 28 for minute 2.

2.4.6 Discussion

Our results provide evidence that the click detection task can uncover the online process of word segmentation from a continuous speech stream. In particular, our study shows that a significant RT difference emerges between the clicks located at the boundaries of words and the clicks located in the interior of the words. Whereas in the first two minutes of exposure the RTs to between- and within-words clicks do not differ, in minutes three and four participants become slower to detect clicks located within words. Congruently with this result, we observe that the group of participants who respond faster to between-words clicks than to their within-words counterparts significantly outnumbers the group showing the opposite trend (24 out of 28, starting from the third minute).

The pattern of results is likely to reflect the segmentation process, because we used a highly controlled acoustic material and two speech streams to counterbalance the target words and part-words. Our results suggest that during exposure to the auditory material, the emergence of word candidates is the main factor determining different RTs associated to both types of clicks. The fact that clicks located within words are associated to longer RTs than clicks located between words could be interpreted as a tendency of participants to expect clicks to occur at the edges of word candidates. The extraction of the transitional probabilities present in the stream (0.33 across word boundaries versus 1.0 within words) might lead participants to have stronger expectations of the following syllable within words, making them less likely to expect extraneous elements such as a click. On the contrary, when the clicks occur between words, weaker expectations are present and participants might be not as surprised. It is important to notice that clicks located between units keep this status throughout the whole stream: at the beginning they are located between syllables, and as participants process statistical information they will still be perceived as being positioned between words. In contrast, clicks located within words are located between syllabic units at the beginning of the exposure, but the extraction of words on the basis of statistics leads them to the interior of word units. We conjecture that processing of the syllables that have become integrated as a putative word will resist interruptions, resulting in longer RTs for clicks within words rather than shorter RTs for clicks between words. The overall increasing trend of RTs might be an indication of fatigue or overload, as participants attended both to the speech material and the clicks.

Asymmetries between edge and interior positions have been also found by Hunt and Aslin (2001), where participants learned statistical information while performing a serial reaction time task. That RTs to clicks located within words are longer also evokes the early findings of click location studies, which showed that clicks positioned in the middle of a syntactic and/or prosodic unit are perceptually moved towards the closest unit boundary (e.g. Fodor et al., 1974). Moreover, such preference for edges of linguistic units is reminiscent of the work on rule generalization by Endress, Scholl, and Mehler (2005). They showed that after a familiarization phase, word structures involving adjacent repetitions were correctly generalized only when repetitions occurred in word edges, and they argued that spontaneous rule generalization happened in this case because of a privileged perceptual saliency of word edge positions. However, it is an open issue for future research to determine whether the word edges induced by statistical learning share a similar perceptual status with the edges that are physically present in the auditory stimuli.

In addition, our findings provide the earliest evidence so far of segmentation with adult participants and behavioral methods, for a significant asymmetry between clicks is already present at the third minute. Our pattern of results is comparable to the ones obtained with a variety of brain imaging techniques (e.g. Abla et al., 2008; Buiatti et al., 2009; Turk-Browne et al., 2009), showing that behavioral methods are able to tap also the initial stages of statistical learning. Moreover, the exposure required by our participants is close to the one used in experiments with young infants for similar types of speech streams (usually two or three minutes). This fact shows that the testing method plays a central role in our ability to assess word segmentation, suggesting also that adults' statistical abilities are not less sharp than infants'.

To the best of our knowledge, our study is the first to use a click detection task with synthesized speech stimuli. This manipulation provided a strong control of acoustics such that RTs to clicks could reveal online processing. Our results support click detection as a method sensitive also to structure (word structure in our case) rather than just to acoustic cues of speech (cf. Cutler et al., 1993).

2.5 Experiment 2
Electrophysiological correlates of segmentation

Looking for a better understanding of the processes underlying segmentation, we now turn to explore neural correlates of segmentation in adult participants. We use a task that is as similar as possible to the classical segmentation experiments and, as in Experiment 1, we focus on the time evolution of selected measures.

Event-related potentials have been used already by several groups to investigate segmentation (e.g. Cunillera, Toro, Sebastián-Gallés, & Rodríguez-Fornells, 2006; De Diego Balaguer et al., 2007; Sanders et al., 2002), but oscillatory brain activity has been less studied. Buiatti et al. (2009) inspected oscillatory activity at specific frequencies related to the structure of their speech material: those corresponding to the duration of monosyllables, bisyllables, and trisyllables (all words in their experiment were trisyllabic). However, these authors did not explore oscillatory activity in traditional frequency bands of EEG.

Oscillatory activity provides naturally an online measure of segmentation, because spectral power can be computed continuously along the EEG signal (unlike event-related potentials). We profit from this continuous nature of oscillatory activity to construct a time course of neural activity during exposure to an artificial speech stream.

We pursued two other main goals in designing an electroencephalographic study of segmentation:

1. The amount of exposure should be comparable to that infants require for successful behavioral segmentation.

2. The experimental design should be easily transferable to infant participants.

Several groups have shown that infants segment succesfully a speech stream after about 2 or 3 minutes of exposure (e.g. Johnson & Tyler, 2009; Saffran, Aslin, & Newport, 1996), in line with the results of Experiment 1. But in order to have a design transferable to test young infants, overt responses to events occurring during the speech stream—such as clicks—were avoided.

We thus decided to focus on oscillatory activity induced by mere (attentive) perception of a word-structured speech stream. Stream length was set to 3 minutes. Following previous works (e.g. Buiatti et al., 2009; Cunillera et al., 2009; McNealy et al., 2006), we included also a stream composed by randomly concatenated syllables, as a baseline condition to which the word-structured stream can be compared.

2.5.1 Participants

Sixteen adults (6 women and 10 men, aged 25 ± 5 years, range 18–39) took part in this study. None of them had neurological, hearing, or language-related deficits. Strength of handedness and lateralization was assessed with the 12-item Edinburgh Inventory (Oldfield, 1971). All participants were right-handed (average index 84%, $SD = 16\%$, range 44%–100%), thirteen of them having indexes over 75%. They were all native speakers of Italian, recruited from the city of Trieste.

2.5.2 Stimuli

Each participant listened to two speech streams, one composed by four bisyllabic words[6] (hereafter, WORD condition) and one composed by randomly ordered syllables (hereafter, RAND condition). The streams, different for each participant, were created using the MBROLA synthesizer with the it4 diphone database. All syllables consisted in a consonant followed by a vowel, with flat intonation and pitch set to 200 Hz. Phoneme duration was 120 ms. Each word was repeated 96 times, for a total stream duration of approximately 184 s.

Streams in the WORD condition were created by concatenating four bisyllabic nonce words in randomized order, with the restriction that no word repeated twice in a row. This way, transitional probabilities between syllables were equal to 1.0 within words, and 0.33 across words. Streams in the RAND condition were created by concatenating eight syllables in randomized order, with the only restriction that no syllable repeated twice in a row. This way, transitional probabilities between syllables were constant and equal to 0.14. See Table 2.1 for examples of word and syllable sets.

All streams were faded in and out for 5 s, to avoid cues to segmentation in addition to transitional probabilities.

[6] The number of syllables per word was reduced with respect to Experiment 1, in order to provide participants with more word exemplars per minute. The number of words was kept the same to assure that a short familiarization would be enough for successful behavioral segmentation.

Participant	WORD condition				RAND condition							
5	puda	nogi	koba	tumi	ba	ki	mu	ne	ta	pi	gu	de
11	geta	poku	noba	demu	ku	ba	gi	mo	tu	da	ni	po
13	nedu	kogi	mopu	tebi	du	pi	ke	ga	mu	ti	ne	ba

Table 2.1: Examples of word and syllable sets for three participants.

2.5.3 Procedure

Participants were tested individually in a soundproof Faraday cage. They listened to one stream of each condition (WORD and RAND), in counterbalanced order. Sound was delivered via a loudspeaker located 1 m in front of the participants, at a sound intensity of 70 dB. Participants used a chin rest, and were instructed to minimize head and body movements. Silent movies were shown in a screen during the presentation of the speech streams.

Participants were told that each speech stream represented an excerpt of a different "alien language". They had to listen to these excerpts, and were instructed to extract the words contained in them. There was a self-paced pause before the beginning of each stream.

After listening to the WORD stream, participants answered a two-alternative forced choice test comprised by 16 trials contrasting words and part-words. For instance, a test item for participant 5 (see Table 2.1) could have been to choose whether "puda" (a word) or "giko" (a part-word, formed by concatenating the last syllable of *nogi* with the first of *koba*) is truly a word of the artificial language. Participants also answered a test after the RAND stream, contrasting pairs of randomly selected bisyllables.

2.5.4 Data acquisition

EEG data were collected using a 128-electrode-net system (Geodesic EEG System 200, Electrical Geodesics, Inc.) referenced to the vertex. EEG signal was bandpass filtered by hardware between 0.1 and 100 Hz, and digitalized at 250 Hz. Electrode impedance was kept below 50 kΩ for at least 90% of all electrodes for each participant[7].

A digital low-pass filter at 90 Hz and a notch filter at 50 Hz were applied to the raw EEG signal. Given the lack of trial structure in our design, we removed from the EEG signal

[7] Only one participant presented slightly higher impedances, with 83% of electrode impedances below 50 kΩ. The inclusion of this participant did not alter our main results.

eye movements, blinks, and other artifacts using Independent Component Analysis instead of rejecting portions of signal. An automatic classifier (Winkler, Haufe, & Tangermann, 2011) computed for each ICA component the probability that it represents an artifact[8]. All ICA components that obtained a probability of 90% or more were discarded after confirmation by visual inspection.

2.5.5 Data processing and analysis

EEG data were referenced offline to the average of all electrodes. We analyzed the time evolution of power in specific frequency bands, searching specifically for bands that tear apart our two conditions.

We computed spectral power in a sliding window of length 2,000 ms shifted in steps of 500 ms. Considering each stream as being 180 s long, this gives a vector of 357 power estimates for each condition. We applied a baseline correction to these vectors by dividing their values by the spectral power in a 2,000 ms window right before stream onset. Corrected power values were then log scaled.

In order to reduce noise, a time course for spectral power was then computed by pooling and averaging the 357 power estimates into 18 bins[9], each one of them representing roughly a window of ten seconds of the speech stream.

This procedure was carried out for each one of the following frequency bands: delta (1–3 Hz), theta (3.5–7.5 Hz), alpha (8–12 Hz), and beta (13–20 Hz). Additionally, gamma band

[8] This classifier is based on six features such as average log power of alpha band and the exponent λ resulting of fitting an inverse power function $f^{-\lambda}$ to the log spectrum f of a component. The classifier computes all six features $\Phi = (\varphi_1, \ldots, \varphi_6)$ and a linear function of them $L(\Phi) = \sum_i w_i \varphi_i - b$ (the parameters w_i and b have been estimated from training data). $L(\Phi) \geq 0$ is associated to components which are more likely artifacts. In our application of this method we used only five of these features, as it was not possible to compute the current density norm of estimated sources for our electrode layout.

[9] The first 17 bins consisted of the average of 20 power values, whereas the last bin consisted of the average of the remaining 17 values.

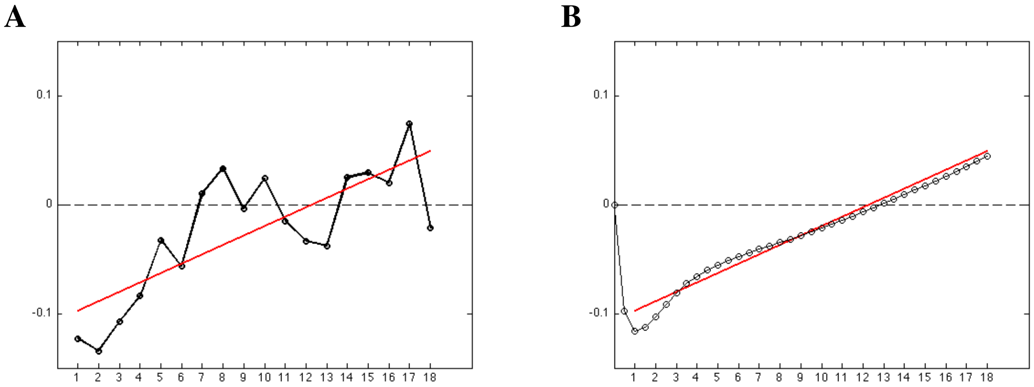

Figure 2.3: Schematic representation of the results and interpretation of the regression parameters. Intercept and slope estimate an initial change and drift in spectral power, respectively. **A.** Simulated spectral data with its regression line. **B.** Idealized trajectory of spectral power, given the same regression line as in A. Spectral power starts form zero (baseline level), presents a fast change given by the intercept, and then follows a drift given by the regression slope.

(20–90 Hz) was divided into seven subbands of 10 Hz[10].

The time courses of 18 power estimates per frequency band and condition were then explored by means of robust linear regressions. As depicted in Figure 2.3, the parameters of these regressions can be interpreted as follows:

The intercept represents the magnitude of the initial change in spectral power with respect to the baseline value.

The slope represents the drift along time induced by the stimulation in spectral power.

We also explored differences between the WORD and RAND conditions in terms of both intercept and slope, by subtracting the power vectors associated to both conditions and applying another robust regression. Additionally, as a means to evaluate the robustness of our

[10] The subbands corresponding to 40–50 and 50–60 Hz partially overlapped with the notch filter applied at 50 Hz. This overlap is relatively small compared to the width of the subbands, and hence we assume that results in these subbands will not be greatly affected. Nonetheless, any influence on the data should affect equally both experimental conditions.

results across subjects, we computed the aforementioned regressions for each subject separately. Statistical assessment of these individual regressions were done by pooling the 16 values of intercept and slope (one per participant), and comparing them against zero with one-sample t-tests.

For all analyses, given that peripheral electrodes tend to present higher levels of noise, we considered the signal of the 92 non-peripheral electrodes.

2.5.6 Results

Behavioral results

The recognition test delivered after the presentation of the WORD stream showed an accuracy of 81% ($SD = 23\%$, range 38%–100%). Seven out of 16 participants responded correctly all test items.

The order in which the two streams (WORD and RAND) were presented did not affect accuracy in the behavioral test ($t(14) < 1$, $p = .88$). The group listening first to the WORD stream responded correctly 80% of test items ($SD = 25\%$, range 38%–100%), whereas the group who listened first to the RAND stream scored 82% ($SD = 22\%$, range 44%–100%). This independence of previous exposure is in line with findings by Franco, Cleeremans, and Destrebecqz (2011), who showed that participants can keep track of statistical information separately for two artificial languages even when their syllable sets overlap.

Time course for each condition separately

We start our time course analysis by looking separately at each condition, by averaging both across subjects and across the 92 non-peripheral electrodes. This should be taken only as an indication of possible effects because it obscures the variability across participants (which can

Frequency band	**Intercept**		**Slope**	
Delta (1–3 Hz)	-0.20	$p < .0001$	0.004	$p = .0021$
Theta (3.5–7.5 Hz)	-0.11	$p < .0001$	0.005	$p < .0001$
Alpha (8–12 Hz)	-0.16	$p < .0001$	0.005	$p = .002$
Beta (13–20 Hz)	-0.16	$p < .0001$	0.004	$p = .008$
Gamma (20–30 Hz)	-0.18	$p < .0001$	0.006	$p < .0001$
Gamma (30–40 Hz)	-0.17	$p < .0001$	0.0007	$p = .53$
Gamma (40–50 Hz)	-0.20	$p < .0001$	0.0009	$p = .50$
Gamma (50–60 Hz)	-0.20	$p < .0001$	0.0006	$p = .65$
Gamma (60–70 Hz)	-0.15	$p < .0001$	-0.0002	$p = .92$
Gamma (70–80 Hz)	-0.19	$p < .0001$	-0.001	$p = .46$
Gamma (80–90 Hz)	-0.16	$p < .0001$	-0.0006	$p = .65$

Table 2.2: Results of the regression analysis for the average curves of spectral power in the WORD condition.

be important in the case of electrophysiological measurements).

The statistics for the WORD condition are presented in Table 2.2. It is evident that attending to the WORD stream elicited a significant reduction in spectral power across all frequency bands, as shown by the negative intercepts. This power reduction was accompanied by significant increases over time (slopes) in the frequency bands up to 30 Hz.

A similar analysis for the RAND condition, presented in Table 2.3, shows again a reduction in power across all frequency bands. This time, however, this reduction was not significant for theta band.

Power variations over time showed a significant linear trend only in alpha, beta, and low gamma band (20–30 Hz).

The time courses for both WORD and RAND conditions, for each frequency band, are depicted in Figures 2.4 and 2.5.

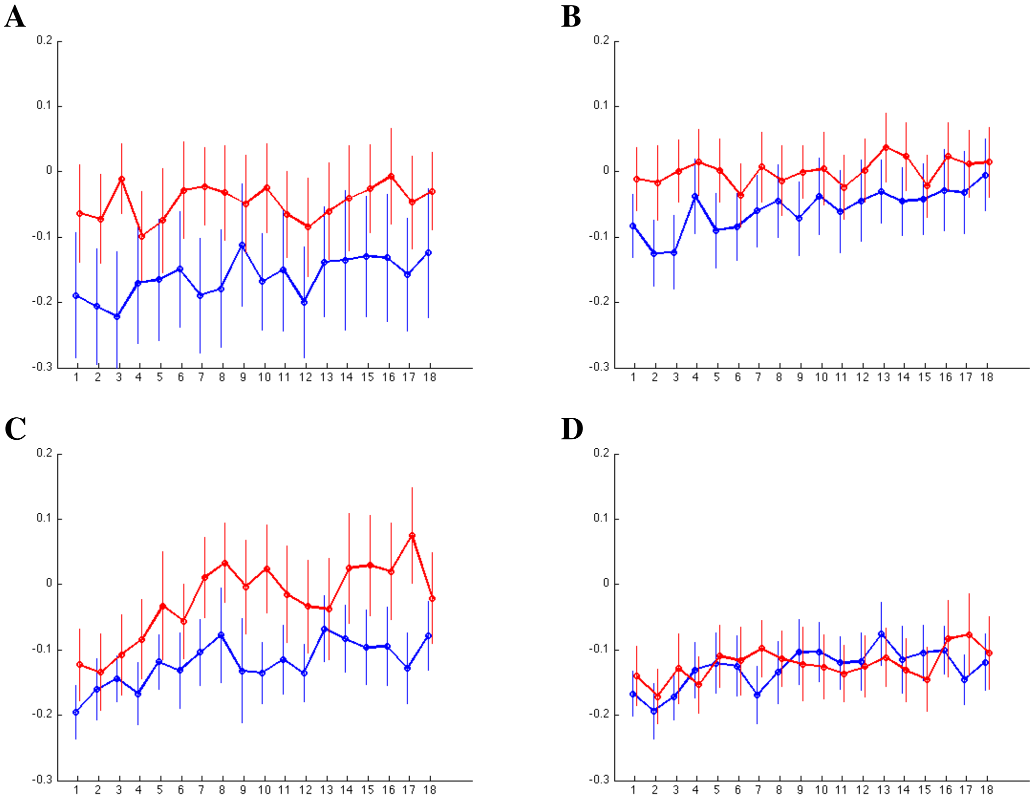

Figure 2.4: Time courses of spectral power evolution in time for the low frequency bands. **A.** Delta (1–3 Hz). **B.** Theta (3.5–7.5 Hz). **C.** Alpha (8–12 Hz). **D.** Beta (13–20 Hz). Blue lines represent the WORD condition, and red lines the RAND condition. Vertical bars depict standard errors.

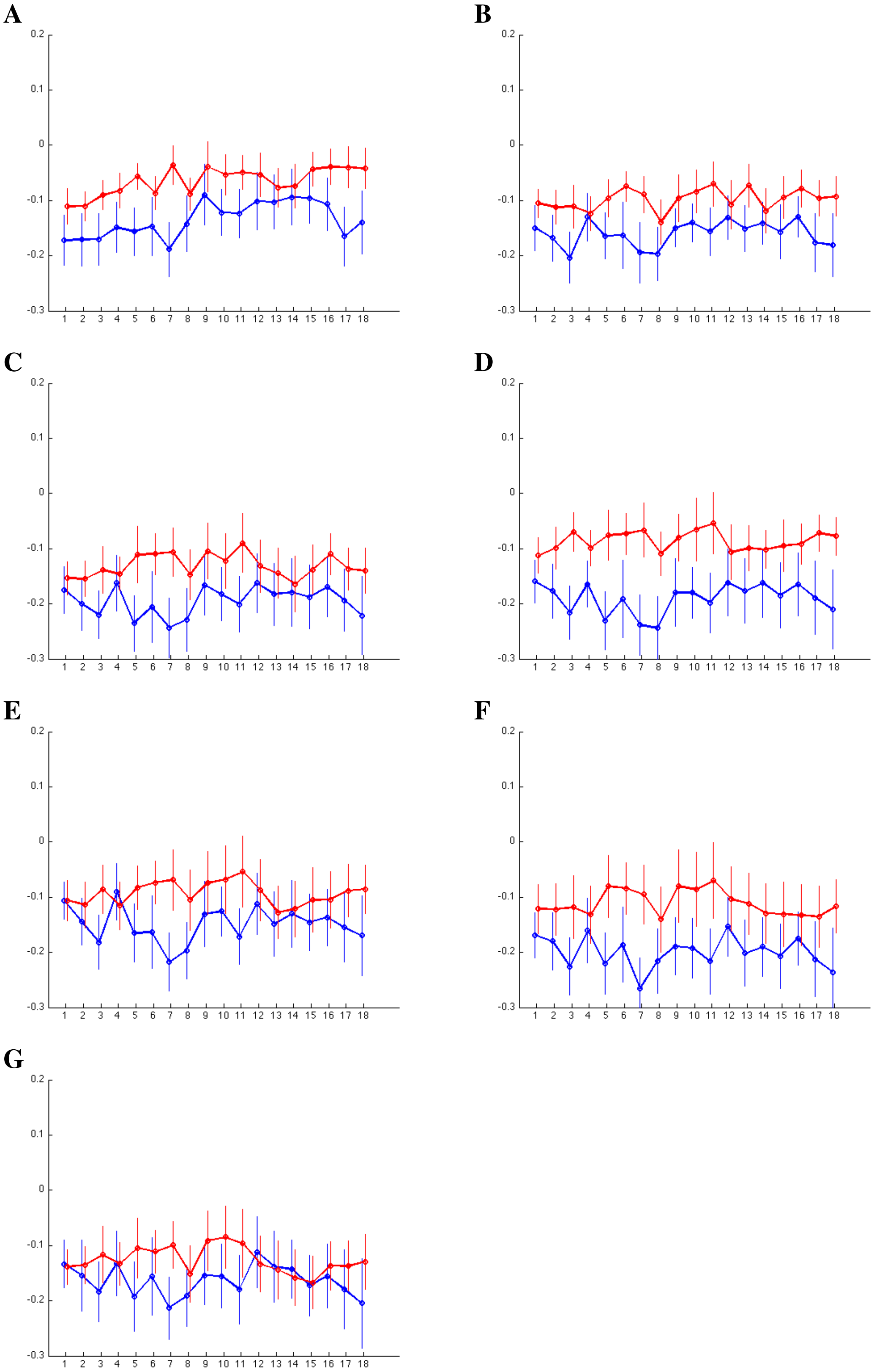

Figure 2.5: Time courses of spectral power evolution in time for gamma subbands. **A.** 20–30 Hz. **B.** 30–40 Hz. **C.** 40–50 Hz. **D.** 50–60 Hz. **E.** 60–70 Hz. **F.** 70–80 Hz. **G.** 80–90 Hz. Blue lines represent the WORD condition, and red lines the RAND condition. Vertical bars depict standard errors.

Frequency band	Intercept		Slope	
Delta (1–3 Hz)	-0.06	$p = .0003$	0.002	$p = .22$
Theta (3.5–7.5 Hz)	-0.01	$p = .25$	0.002	$p = .11$
Alpha (8–12 Hz)	-0.11	$p = .0001$	0.009	$p = .0004$
Beta (13–20 Hz)	-0.14	$p < .0001$	0.002	$p = .029$
Gamma (20–30 Hz)	-0.10	$p < .0001$	0.004	$p = .0014$
Gamma (30–40 Hz)	-0.11	$p < .0001$	0.0012	$p = .21$
Gamma (40–50 Hz)	-0.13	$p < .0001$	0.0003	$p = .80$
Gamma (50–60 Hz)	-0.09	$p < .0001$	0.0003	$p = .75$
Gamma (60–70 Hz)	-0.09	$p < .0001$	≈ 0	$p = .97$
Gamma (70–80 Hz)	-0.10	$p < .0001$	-0.0008	$p = .48$
Gamma (80–90 Hz)	-0.11	$p < .0001$	-0.0012	$p = .30$

Table 2.3: Results of the regression analysis for the average curves of spectral power in the RAND condition.

Differences between conditions

We now explore the differences between WORD and RAND. Intercepts and slopes for the subtraction of both conditions and all frequency bands are presented in Table 2.4.

The presence of negative intercepts in all frequency bands indicates that the reductions in spectral power were larger for the WORD condition than for the RAND condition, although significance varied widely across frequency bands.

Differences in slope, that is in the evolution of spectral power during the exposure to the streams, displayed a more specific pattern: significant differences are only reached for theta and alpha bands.

We conduct now a similar but more stringent analysis, considering inter-participant variability. As described in Section 2.5.5, we recomputed the linear regressions separately for each participant (average signal of the 92 non-peripheral electrodes), and compared the samples of 16 intercept and 16 slope values against zero through t-tests. Tables 2.5 and 2.6 summarize average values, standard deviations, and p values for intercepts and slopes, respectively. As anticipated, results of this analysis are more strict, with significant results in only one fre-

Frequency band	**Intercept**		**Slope**	
Delta (1–3 Hz)	-0.13	$p < .0001$	0.002	$p = .23$
Theta (3.5–7.5 Hz)	-0.09	$p < .0001$	0.003	$p = .007$
Alpha (8–12 Hz)	-0.06	$p = .016$	-0.005	$p = .024$
Beta (13–20 Hz)	-0.02	$p = .27$	0.002	$p = .34$
Gamma (20–30 Hz)	-0.07	$p = .0001$	0.0008	$p = .52$
Gamma (30–40 Hz)	-0.06	$p = .001$	-0.0004	$p = .79$
Gamma (40–50 Hz)	-0.07	$p = .004$	0.0005	$p = .79$
Gamma (50–60 Hz)	-0.11	$p = .0001$	0.0004	$p = .86$
Gamma (60–70 Hz)	-0.06	$p = .035$	≈ 0	$p = .99$
Gamma (70–80 Hz)	-0.09	$p = .0008$	≈ 0	$p = .99$
Gamma (80–90 Hz)	-0.04	$p = .06$	0.0007	$p = .74$

Table 2.4: Results of the regression analysis for the difference of the average curves of spectral power for the WORD and RAND conditions.

quency band for each parameter. The WORD condition elicited a larger reduction in spectral power in gamma band (50–60 Hz) than the RAND condition, and it presented a larger slope in theta band (3.5–7.5 Hz).

Correlation with behavior

We evaluated the correlation between the two effects found in the previous section and the results of the behavioral test.

Slopes of the evolution in time of theta power were negatively correlated to accuracy in the behavioral test ($\rho = -.40$), although this value did not reach statistical significance ($p = .12$). Similarly, we observed a non-significant negative correlation between intercepts in gamma (50–60 Hz) band and behavioral results ($\rho = -.46$, $p = .08$).

See Figure 2.6 for scatterplots of both effects, and Section 2.5.7 for an interpretation of their direction.

Frequency band	Intercept	t-statistic	
Delta (1–3 Hz)	−0.15(0.37)	−1.60	$p = .13$
Theta (3.5–7.5 Hz)	−0.09(0.22)	−1.62	$p = .13$
Alpha (8–12 Hz)	−0.05(0.24)	< 1	$p = .40$
Beta (13–20 Hz)	−0.02(0.18)	< 1	$p = .70$
Gamma (20–30 Hz)	−0.07(0.23)	−1.31	$p = .21$
Gamma (30–40 Hz)	−0.05(0.20)	< 1	$p = .34$
Gamma (40–50 Hz)	−0.07(0.16)	−1.68	$p = .11$
Gamma (50–60 Hz)	−0.11(0.19)	−2.24	$p = .041$
Gamma (60–70 Hz)	−0.04(0.22)	< 1	$p = .46$
Gamma (70–80 Hz)	−0.09(0.25)	−1.46	$p = .17$
Gamma (80–90 Hz)	−0.04(0.18)	< 1	$p = .35$

Table 2.5: Average differences across participants (standard deviations in parentheses) between WORD and RAND conditions for the estimated intercepts of the per-subject regression analysis. These values were submitted to a one-sample t-test (15 degrees of freedom), whose t-statistic and p value are also presented.

Frequency band	Slope	t-statistic	
Delta (1–3 Hz)	0.003 (0.01)	1.33	$p = .20$
Theta (3.5–7.5 Hz)	0.003 (0.005)	2.69	$p = .017$
Alpha (8–12 Hz)	−0.004 (0.02)	< 1	$p = .42$
Beta (13–20 Hz)	0.001 (0.009)	< 1	$p = .59$
Gamma (20–30 Hz)	0.0005(0.008)	< 1	$p = .80$
Gamma (30–40 Hz)	−0.0009(0.006)	< 1	$p = .51$
Gamma (40–50 Hz)	0.0005(0.008)	< 1	$p = .81$
Gamma (50–60 Hz)	0.0005(0.007)	< 1	$p = .78$
Gamma (60–70 Hz)	−0.0012(0.01)	< 1	$p = .63$
Gamma (70–80 Hz)	0.0004(0.008)	< 1	$p = .85$
Gamma (80–90 Hz)	0.001 (0.007)	< 1	$p = .57$

Table 2.6: Average differences across participants (standard deviations in parentheses) between WORD and RAND conditions for the estimated slopes of the per-subject regression analysis. These values were submitted to a one-sample t-test (15 degrees of freedom), whose t-statistic and p value are also presented.

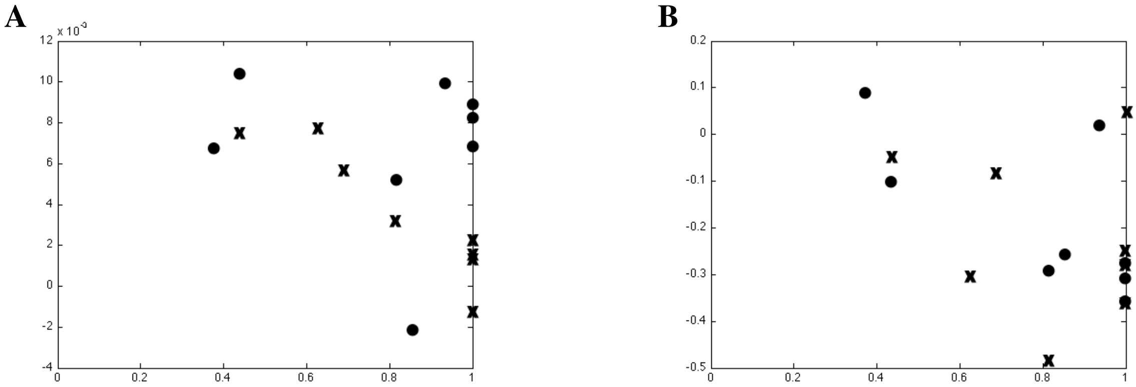

Figure 2.6: Scatterplots of behavioral scores in the test after the WORD condition (horizontal axes) versus neural scores (vertical axes). **A.** Slopes of the increase in theta power in the vertical axis. **B.** Intercepts of gamma (50–60 Hz) power in the vertical axis. Circles mark participants who listened to the WORD stream first, whereas crosses mark participants who listened to the RAND stream first.

Topography of power changes

Up to now, we have identified two frequency bands that are involved in the process of attending to an artificial speech stream containing words, as opposed to attending to a stream lacking this structure. These results were discovered from the examination of the average signal of all non-peripheral electrodes.

We now turn to explore the scalp regions more involved in these effects. First, observe that the increases in theta power for the WORD condition seem to be spread all over the scalp. This is evident from the fact that 86% ($SD = 23\%$) of the 92 non-peripheral electrodes display positive slopes. The analog measure for the RAND condition reaches 62% ($SD = 30\%$), significantly lower (paired t-test, $t(15) = 2.29$, $p = .037$). As for the differences in intercept, both conditions display a similar number of non-peripheral electrodes presenting a reduction in gamma (50–60 Hz) power with respect to baseline (WORD: 85%, $SD = 17\%$, RAND: 80%, $SD = 20\%$). However, 66% ($SD = 25\%$) of the electrodes display larger reductions in the WORD condition ($t(15) = 2.67$, $p = .017$).

Figure 2.7 displays topographic maps depicting average slopes for theta band and average intercepts for gamma band. As we have pointed out, both effects look similar between conditions but WORD presents larger magnitudes. To look for significant differences between conditions in the topographic maps, we make use of the clustering method presented by Maris and Oostenveld (2007), whose adaptation to our case we describe briefly in the following paragraphs.

The clustering method is a statistical test that allows to conduct multiple comparisons without the need to recur to stringent correction procedures such as Bonferroni's. The main idea behind this method is to use information about the spatial distance between electrodes in order to group neighboring active electrodes in larger units or *clusters*. Then, a permutation test is conducted at the cluster level, converting a problem of multiple comparisons concerning electrodes into a single comparison concerning clusters.

To define a matrix of neighboring electrodes, we computed the distribution of inter-electrode distances of a standardized EGI net (i.e. the distribution of the distance between all pairs of electrodes; see Figure 2.8). Visual inspection shows that the first peak in this histogram finishes in a local minimum at about 3.6. Thus we assumed two electrodes to be neighbors if their distance is smaller than this value.

The clustering method works as follows: Given a threshold value $\xi > 0$, we consider all the electrodes whose t-statistic for the contrast between conditions is larger (in absolute value) than ξ. Neighboring electrodes are then pooled together in connected clusters[11], and a cluster-level statistic is computed as the sum of the t-statistics of the electrodes composing each cluster. The largest cluster-level statistic (in absolute value) is considered the statistic T of the whole array, and its associated cluster is the localization to be statistically assessed. This assessment is then carried out by means of a *permutation test*: the process of calculating T is

[11] In each cluster, the t-statistics for all channels must have the same sign (i.e., all positive or all negative).

repeated after randomly swapping the two conditions for each subject, computing a histogram for T_{rand} against which the true value of T is compared[12]. p values for such test are computed as the percentage of area in the histogram that are larger than the true T. The resulting p value is then compared to the desired α level.

There is no value for ξ that is valid for all datasets. ξ must be chosen based on the distribution of t-values to be examined[13]. For our data, we used thresholds $\xi = 1.8$ and $\xi = 1.9$ for the theta and gamma band contrasts, respectively.

The theta band contrast reveals a significant cluster ($p = .045$) comprising 34 electrodes spread across frontal, right-frontal, and left-central areas. The gamma band contrast also shows a significant cluster ($p = .013$) comprising 34 electrodes, but distributed across posterior and right areas. Figure 2.9 presents the scalp distributions of these clusters.

Analysis of the first stream

As a final analysis, we recalculated the statistical tests for slopes in theta band and intercepts in gamma band considering only the first stream to which each participant was exposed. We thus reduce our within-subjects design with $N = 16$ to a between-subjects design with $N = 8$ in each group. This modification reduces the statistical power of our contrasts, but avoids possible confounds due to order effects in the presentation of the two conditions.

The theta band slope contrast turned significant in this restricted setting ($t(14) = 2.66$, $p = .019$), supported by a substantial increase of the effect size measured by Cohen's d: $d = 0.82$ for the original, within-subjects contrast, compared to $d = 1.12$ for the between-subjects design.

[12] For our analysis, we built this histogram with $1,000$ random reassignments.

[13] ξ can be tuned in order to detect larger or smaller clusters: a lower threshold will lead to larger clusters, but clusters that are too large are unlikely to be significant at the α level, according to the permutation test.

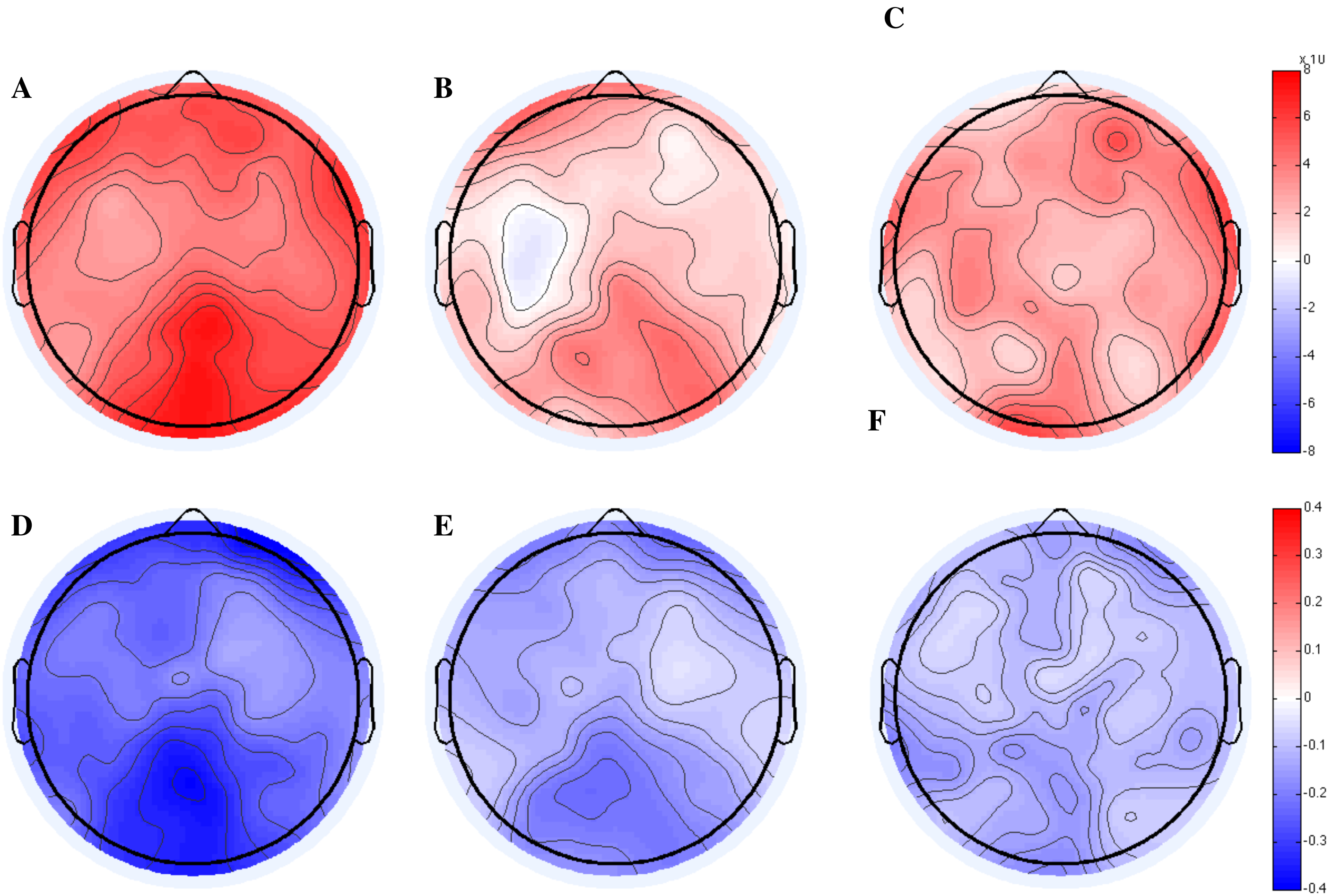

Figure 2.7: Topographic maps showing the scalp distribution of the effects in theta band slope (A–C) and gamma band intercept (D–F). **A.** Slopes for theta power in the WORD condition. **B**. Slopes for theta power in the RAND condition. **C.** Difference between A and B. **D.** Intercepts for gamma power in the WORD condition. **E.** Intercepts for gamma power in the RAND condition. **F.** Difference between D and E.

As for the gamma band intercept contrast, it was not significant ($t(14) = -1.32$, $p = .21$). Since in this case the effect size for the between-subjects contrast ($d = -0.64$) was slightly larger than the original within-subjects effect size ($d = -0.57$), we interpret the lack of significance as due to a reduction in statistical power because of the change in design.

2.5.7 Discussion

We have explored the neural correlates of segmentation of continuous speech and their evolution in time. Even if several groups have approached this problem before (e.g. Abla et al., 2008; Buiatti et al., 2009; Cunillera et al., 2006; De Diego Balaguer et al., 2007; Sanders et al., 2002; Teinonen et al., 2009), the neural correlates of segmentation in oscillatory activity in traditional frequency bands had not been considered in depth.

Our work partly aimed to fill this gap in the literature, by exploring the time course of spectral power during exposure to speech streams composed by words or random syllables. Results demonstrate the presence of two effects:

- A robust increase in power in theta band (3.5–7.5 Hz) during exposure to a continuous speech stream containing words. This increase in time differs significantly from what occurs when the word structure is absent.

- A larger tonic reduction in spectral power in gamma band (50–60 Hz) for the speech stream containing words.

These results hold both at the average level and in a further analysis that considers subject-to-subject variability. It is worth noticing that 14 out of 16 participants presented positive slopes for theta power in the WORD condition ($p = .004$, binomial test), whereas only 8 out of 16 did so in the RAND condition.

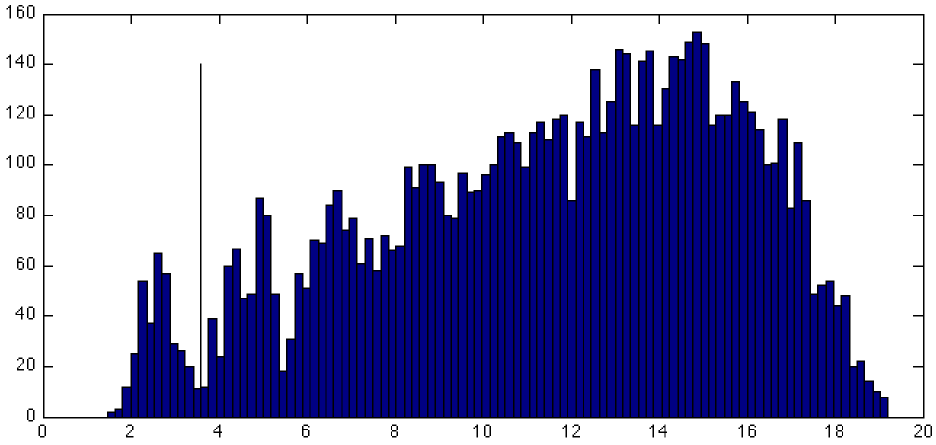

Figure 2.8: Histogram of inter-electrode distances in a standardized EGI net with 128 electrodes. The vertical line marks the end of the first peak. All pairs of electrodes whose distance was shorter were considered neighbors for the clustering analysis.

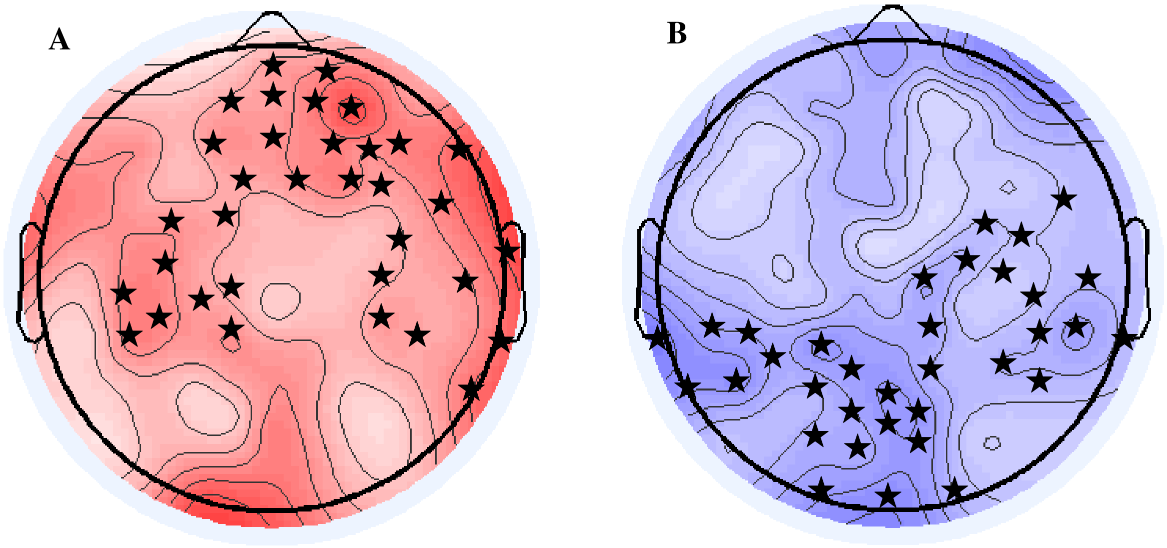

Figure 2.9: Results of the clustering analysis for the effects in theta band (A) and gamma band (B). Stars mark the location of the electrodes belonging to the clusters differing between WORD and RAND conditions (clustering analysis with $\alpha = 5\%$).

Results of the behavioral data showed a high number of participants performing at ceiling level (7 out of 16), suggesting that the selection of four bisyllabic words for the WORD stream was too easy. This decision was made in order to maximize the number of words presented per minute in the WORD stream, to make sure that successful segmentation would occur. The downside is that words recur with high frequency, every two seconds approximately. This ceiling effect might have also impaired the correlation between neural and behavioral results. Future work should explore this possibility by using more complex material.

The role of brain oscillations in segmentation

Few studies have investigated brain oscillations during the segmentation of artificial speech. As we have already mentioned, Buiatti et al. (2009) conducted a study similar to ours but focusing on the extraction of a non-adjacent dependency. Their analysis profited from the rhythmic regularity given by the fact that all words in their material were trisyllabic, by looking at spectral power at the specific frequencies corresponding to the duration of one-, two-, and tri-syllabic units. Their ingenious contrast of different kinds of speech streams allowed them to obtain results that went beyond simple detection of rhythmicity. Nonetheless, such analysis was closely tailored for the case of artificial languages in which all words have the same length, and it is unclear how it could be extended to the general case. Our analysis focused instead on traditional frequency bands, providing clear and testable predictions for future studies with simple or complex languages[14].

Before discussing the specific frequency bands that responded differently to streams with and without word structure, we point out that the vast majority of the literature studying brain oscillations refers to oscillatory effects occurring at the millisecond scale. We have

[14] Notice that our results cannot be reduced to the same rhythmic explanation. The duration of each word was 480 ms. This corresponds to a frequency of 2.083 Hz, which belongs to delta band. In this frequency band, however, we did not detect any significant difference between conditions.

deliberately overlooked those variations in order to focus on the long term changes, meaning that the comparison between our results and the literature must be considered as tentative.

Activity in theta band has been linked to several processes, such as response inhibition (e.g. Kirmizi-Alsan et al., 2006), language translation (Grabner, Brunner, Leeb, Neuper, & Pfurtscheller, 2007), processing of semantic relationships (e.g. Bastiaansen, van der Linden, ter Keurs, Dijkstra, & Hagoort, 2005; Hald, Bastiaansen, & Hagoort, 2006; Maguire, Brier, & Ferree, 2010), drowsiness, arousal, and sustained attention (e.g. Huang, Jung, & Makeig, 2009; Lin et al., 2010; Makeig & Jung, 1996), and memory enconding (e.g. Jensen & Tesche, 2002; Klimesch, Doppelmayr, Russegger, & Pachinger, 1996; Mölle, Marshall, Fehm, & Born, 2002; Sederberg, Kahana, Howard, Donner, & Madsen, 2003). From these possibilities, the last two are the most applicable to our current study, and we will discuss them in the following paragraphs.

In studies of drowsiness, arousal, and sustained attention, EEG is recorded typically while participants have to monitor an auditory signal (Makeig & Jung, 1996) or correct the direction of a virtual car in a simulated driving task (Huang et al., 2009; Lin et al., 2010). These studies have found a power increase in alpha and theta bands associated with an increase in the probability of missing a target or with longer reaction times. This increase is suppressed in the case of non-missed targets. Given the monotonicity of our experimental material, drowsiness seems a possible explanation to the sustained increase in theta power we observed. However, this alternative does not account for the difference between the WORD and RAND conditions. Additionally, the observed scalp distribution of drowsiness-related effects is posterior, in disagreement with our data.

Memory tasks and encoding of new information have also been associated to spectral power in theta band. Jensen and Tesche (2002) observed increases in theta power proportional to the number of items retained in working memory. Sederberg et al. (2003) observed that the probability of recalling an item was predicted by the increase of theta power during encoding. Mölle et al. (2002, see also Klimesch et al., 1996) proposed that successful intentional encoding of paired words and faces elicited a combined effect of theta synchronization and alpha desynchronization. Interestingly, the scalp distribution of this type of theta activity tends to be frontal, and some groups like Sederberg et al. (2003) have reported the involvement of not only frontal but also right temporal areas (compare to Figure 2.9).

These lines of research lead us to hypothesize a dual role of theta oscillations in segmentation. The initial dip in theta power with respect to baseline observable in Figure 2.4B (which is significant: $t(15) = 2.16$, $p = .047$) may be due to extra attentional resources allocated to the WORD stream, whereas the later upward drift (the significantly positive slope) may reflect incremental strengthening of the memory traces of the segmented words. This hypothesis, however, would imply that larger slopes are associated to higher accuracies in the behavioral test, which turned not to be the case. A discussion of the direction of the correlation between our neural and behavioral results can be found in the next section.

As with theta oscillations, brain activity in the gamma frequency band has also been linked to memory processes (e.g. Kaiser & Lutzenberger, 2005; Sederberg et al., 2007), with some studies observing combined effects in gamma and theta frequency bands (Lisman, 2010; Nyhus & Curran, 2010). Gamma band has been additionally associated to selective attention (e.g. Fell, Fernández, Klaver, Elger, & Fries, 2003), muscular activity (Pope, Fitzgibbon, Lewis, Whitham, & Willoughby, 2009), and to the perception of the native language in young infants (Peña et al., 2010). However, gamma activity typically presents significant increases in response to relevant stimulation (e.g. Fitzgibbon, Pope, Mackenzie, Clark, & Willoughby, 2004), and very few studies have focused on desynchronization in this frequency range. Ihara et al. (2003) observed desynchronization in frequencies below 50 Hz in a syntactic processing task. Notably, this desynchronization localized to several language-related areas in the left hemisphere. Gómez, Vaquero, López-Mendoza, González-Rosa, and Vásquez-Marrufo (2004) instead studied oscillatory activity in the expectancy period that follows a visual cue and precedes the presentation of a stimulus. These authors observed a generalized reduction in spectral power with respect to baseline, in frequencies ranging up to about 40 Hz and localized frontally for the frequencies between 30 and 40 Hz. This reduction would serve the purpose of maximizing the sensitivity of the system for the incoming stimulus[15]. None of these observations, however, seems to be in line with our results, which show a reduction in power for both conditions, localized to posterior and right-central electrodes. This reduction was larger for the WORD stream, indicating that the relevant process may be neural desynchronization with respect to baseline activity. Future work should assess the replicability of these differences in gamma oscillations, and explore its role in depth.

Direction of the correlation with behavior

The negative sign of the correlations between our neural results and accuracy in the behavioral test for the condition seems puzzling at a first glance. However, they can be taken as indications to which subprocesses are the relevant ones for successful segmentation.

In the case of gamma (50–60 Hz) activity, we observed a significant intercept with a non-significant slope. If, as previously conjectured, this is interpreted as neural desynchronization, then the magnitude of desynchronization is actually inversely related to the intercept parameter. More specifically, since our data was log-scaled and thus the baseline value is zero, the initial reduction in spectral power is equal to the intercept with its sign changed. This changes the direction of the correlation, rendering its sign positive and supporting the interpretation of neural desynchronization as the crucial process.

[15] In line with this, Hanslmayr, Staudigl, and Fellner (2012) have recently proposed an “information via desynchronization” hypothesis, suggesting that reductions in oscillatory power are important processes for the encoding and retrieval of memory traces.

A similar case can be done about theta activity, where we found a significant positive slope representing a possibly dual process: an initial desynchronization and a subsequent slow re-synchronization. Since the evolution of theta power starts at zero (baseline level) and comes back to about the same level (see Figure 2.4B), the magnitude of these two components, desynchronization and re-synchronization, are coupled. The negative sign of the correlation between behavioral performance and the re-synchronization slopes thus suggests that the relevant process guiding segmentation is the initial desynchronization. Indeed, the magnitude of this initial desynchronization (measured again by the intercept of the regression with its sign changed)[16] correlates positively with the behavioral results ($\rho = .38$), with a significance comparable to the originally computed values ($p = .14$).

Limitations

Our study has several limitations that should be addressed by future research. Maybe the most important of them: our study lacked trial structure. In other words, our design consisted in two experimental conditions, each one of them with a single trial lasting for three minutes. Some shortcomings of such designs are well known: lower signal-to-noise ratios and higher chances of artifacts contaminating the signal. In spite of these limitations, this kind of "single trial" designs have already been used in the literature, especially when the object of study is the evolution across time of a continuous process (e.g. Buiatti et al., 2009; Mikutta, Altorfer, Strik, & Koenig, 2012). Other researchers have successfully used ERPs computed for words and part-words during time windows of exposure (e.g. Abla et al., 2008; Abla & Okanoya, 2009), but with material that is acoustically simpler than continuous speech (e.g. pure tones or visual shapes). Continuous speech lacks long periods of stable acoustic information (even

[16] The magnitude of this desynchronization was significant for the WORD condition ($t(15) = 2.16$, $p = .047$), but not for RAND ($t(15) < 1$, $p = .80$).

vowels are affected by coarticulation with the following consonant), difficulting the selection of an appropriate baseline period.

Although we checked carefully all the ICA components from each participant to discard artifacts, it is possible that some of them went undetected. Still, we can assume that the influence of undetected transient artifacts in our results is not strong, because our analysis included both averaging in time and the use of robust regression (which is resistant to outliers).

2.6 Chapter discussion

In this chapter, we presented two experiments investigating the time course of word segmentation from a continuous speech stream.

Experiment 1 demonstrated that the latency to detect clicks can index segmentation of an artificial stream, by measuring the difference in reaction times to clicks located either between or within words. Results showed a steady increase in time of the gap between reaction times for the two types of clicks, and that this gap reached significance by the end of the third minute of exposure.

Experiment 2 explored the neural correlates of segmentation by looking at brain oscillations in different frequency bands. We found a robust effect in theta band, characterized by a two-stage process when participants were exposed to a stream containing words:

- A fast initial desynchronization, shown by a significant reduction in spectral power with respect to baseline, and
- a subsequent re-synchronization drift, evidenced by a significant increasing slope during the listening.

2.6.1 The role of attention in word segmentation

In our behavioral investigation in Experiment 1, we did not conduct an offline test of segmentation after exposure to the speech stream. This leaves open the question of whether segmentation indexed by latencies to clicks and a recognition test correlate. The question is far from trivial, because click monitoring takes attentional resources away from the speech stream.

Several authors (e.g. Perruchet & Pacton, 2006; Toro, Sinnett, & Soto-Faraco, 2005; Turk-Browne, Jungé, & Scholl, 2005) have argued that attention is critical for segmentation according to offline tests, although they lacked tools to determine if this interference affected the segmentation process itself or the conscious recall of segmented sequences. Nonetheless, Saffran, Newport, Aslin, Tunick, and Barrueco (1997) have reported that both children and adults succeed in offline tests even when incidentally exposed to the speech stream, during a simple concurrent task. Click detection can be a rather demanding concurrent task, due to the speeded responses to a stimulus that shares the same modality as speech. This leads us to hypothesize that a click detection task might interfere with offline tests to assess performance.

Perruchet and Vinter (1998) suggested that attending to the speech stream and the statistical regularities therein contained is sufficient to succeed in segmenting it, even if this attention is directed for a very short period. Pacton and Perruchet (2008) took further this argument, arguing that attention is a necessary and sufficient condition for the extraction of these regularities.

2.6.2 Correspondence between neural and behavioral segmentation

Results of Experiment 2 indicate that the segmentation process may be composed of two stages, a first one characterized by directed attention to the speech stream, and a second one related to memory processes.

We observed that the first stage is already able to predict results of the final behavioral test with the same strength as other candidate measures, by showing that the initial dip in theta power tends to correlate positively with behavior. Given that the first data point in the time courses depicted in Figures 2.4 and 2.5 represents just the first 10 seconds of exposure to the stream, this would suggest that segmentation is actually a strikingly quick process. Yet, this conclusion must be taken with a grain of salt because of the simplicity of the material we used, comprising only four bisyllabic words. The total duration of these words is 1,920 ms, implying that after 10 s of exposure each word has been heard about five times. This high recurrence might have been sufficient for participants to segment quickly, devoting the rest of the listening time to commit these words to memory (implicitly or explicitly).

Based on this, we make two predictions for a possible future experiment with a similar procedure but different, more complex, material:

- The initial "attentive" stage, in which spectral power while listening to the WORD stream keeps significantly below zero, will last longer.
- After this initial stage, a slow drift of re-synchronization will take spectral power back to its basal level.

2.6.3 Implications for segmentation models

Our two experiments provide two complementary measures for the time course of segmentation. But whereas in the neural data (Exp. 2) we observed a dual-stage process, in behavioral data (Exp. 1) only the steady separation between the two conditions was present. If indeed, as the neural data insinuates, segmentation occurs quickly and then leaves place to a steady memory enhancement stage, one could argue what is it that the click detection task is indexing. As suggested by the neural data, latencies to respond to click may be determined by the strength of the proto-lexical memory traces corresponding to the artificial words, instead of segmentation *per se*. Still, our neural data does not clarify what would happen when using more complex artificial languages.

The dual-stage model is reminiscent of Perruchet and Vinter's (1998) PARSER model, based on the storage of chunks of syllables in a proto-lexical space. These stored chunks are subsequently strengthened or weakened according to their recurrence in the speech stream. If this neural pattern of results holds for more general material, neural data on the time course of segmentation may provide indirect support to this model. PARSER turns out to be sensitive to transitional probabilities and other distributional properties of the speech stream without encoding them explicitly. The necessity of explicit encoding of statistical information, or more generally how statistical information of the speech stream is encoded and stored in memory, is so far an unaddressed question. Early claims by Aslin et al. (1998) argued that transitional probabilities are the only necessary first-order statistics to segment, although this does not rule out that listeners encode other information. Indeed, adult and infant listeners also seem to encode backward transitional probabilities (Pelucchi et al., 2009; Perruchet & Desaulty, 2008), which may be taken as an indication that sensitivity to transitional probabilities is a by-product of the segmentation process, rather than its driving force.

2.6.4 Final remarks

By looking at the behavioral and neural time courses of speech segmentation, we have explored possible indexes of this process. Latencies to respond to clicks and neural oscillations in theta band emerged as candidates. Theta band presented a dual-stage pattern of results, with a quick initial desynchronization and a subsequent steady re-synchronization. Future work should aim to replicate and extend this investigation, and to test empirically the possible coupling between behavioral and neural results. Such program of research may not only provide a better picture of the evolution in time of speech segmentation, but also to enlighten the nature of the underlying representations.

Chapter 3

Exploring constraints on the primordial syllable

In Section 1.3.1, we reviewed evidence indicating that the basic perceptual unit for speech at birth is the syllable. However, it is unclear how this *primordial syllable* relates to the adult concept of syllable. This distinction is important, since acquiring a language involves a strong component of perceptual learning (see Sections 1.2.1 and 1.2.2 on the acquisition of rhythm and phonetic repertoire, respectively), whereby the perceptual landscape for speech is deeply reshaped during development. The extraction of these primordial syllables from continuous speech is a central task at the first stages of speech processing.

Properties that are common to all languages, called *language universals*, constitute a good starting point for developmental scientists to look for regularities reflecting the initial state of human speech perception.

In this chapter, we will focus on a series of universal phonological regularities explained by the Sonority Sequencing Principle. This principle describes constraints on syllabic structure, implying the existence of a scale of preference among all possible syllables[1] and, therefore, of well- and ill-formed syllables.

We will explore the sensitivity of newborn infants, who lack significant linguistic exposure, to these phonological regularities by means of three speech perception experiments. This will shed light on the question whether the initial state of syllabic perception is also affected by this phonological universal preference.

3.1 Phonological universal preferences

Recall that (see Section 1.1.1) *Universal Grammar* (UG) was a term originally used to denote the set of facts about language structure that are common to all languages. In this sense, UG is just a recollection of common elements across languages. Chomsky's redefinition of this term reified UG as a set of constraints guiding language acquisition, similar to constraints of the early visual system or other perceptual systems (Chomsky, 2005). Thus, UG is assimilated to the human biological endowment for language, and universal properties common to all languages become a consequence of the universality of UG, so that every property that is specified in UG must be true of every language.

Different linguistic proposals for specifying the knowledge about language that is contained in UG vary in many respects, hinting at the complexity of this task (e.g. see Table 2 in Fitch, 2011). This complexity is made evident by typologists, who study variability across human languages, with not few of them rejecting the existence of any universal language property (Evans & Levinson, 2009). The basic reason behind this is that for most properties claimed to be universal, there seems to be at least one language lacking them. One of the most famous recent claims is the one made by Everett (2005), who suggests that the amazonian

[1] By "possible syllable", here we refer to any sequence of phonemes containing a vowel.

language Pirahã lacks recursion.

Criticisms like Evans and Levinson's (2009; see also Dunn, Greenhill, Levinson, & Gray, 2011) triggered a series of reactions in the linguistic and cognitive scientific communities to look for a better, integrated understanding of the origin of properties of language. However, the debate has typically overlooked phonology as a potential source of universal constraints on languages (or at least, spoken languages). In this direction, recent developments in psycholinguistic research regarding possible phonological universals are important to consider. Work by Iris Berent and colleagues (2007; 2008) point specifically to restrictions on syllabic structure. These restrictions are briefly summarized as follows: most speakers of languages that allow consonant clusters at syllable onset agree that *bre* is a possible, although probably inexistent, word. However, the same speakers may be reluctant to accept *rbe* as a possible word, even though they most probably have never experienced either one or the other. Such phonological preference may be a product of linguistic experience, whereby words with a *br-* onset outnumber words with a *rb-* onset across languages. But this preference may also reflect a universal preference for the sound structure of language, that is, a principle comprising UG. As such, this preference should be available to speakers of every language, including those lacking consonant clusters in their syllabic onsets. Berent et al. (2008) made a first step in this direction, testing perception of a variety of syllabic onsets with adult listeners of Korean, a language with a very poor set of onset consonant clusters.

In the following section, we will review the linguistic concept of *sonority*, and how it is used to describe universal phonological preferences such as the one mentioned above.

3.1.1 Sonority

Sonority is a phonological property of speech sounds, roughly correlated to intensity[2] (Parker, 2002). It refers to the difference in pressure between the air internal and external to the oral cavity when a specific phoneme is produced. For instance, whereas vowels (e.g. /a/, /u/) are produced with an open mouth so that no pressure difference is induced, obstruents (e.g. /b/, /f/) require a period of full occlusion and accumulation of air in the mouth, building up the highest pressure difference. All other phonemes lie in between these two extremes, creating a scale or hierarchy. Sonority is simply the reverse of this scale: phonemes that induce little pressure differences are considered *more sonorant* than phonemes that induce high pressure differences. Thus vowels are the *most sonorant* phonemes, and stop consonants are the *least sonorant* ones. The following scheme depicts a simple version of this Sonority Hierarchy[3], with exemplars of each sonority level.

Obstruents		Nasals		Liquids		Glides		Vowels
/b/, /f/	<	/m/, /n/	<	/l/, /r/	<	/w/, /j/	<	/i/, /æ/
bead		*mouse*		*loud*		*wet*		*any*
fee		*nose*		*rope*		*yes*		*ask*

The earliest precursors of scales such as sonority can be traced back to the work of Sanskrit grammarians like Pāṇini (500 B.C.), who grouped the phonemes of Sanskrit into 14 *natural classes* according to their degree of "opening" (vivāra; cf. Parker, 2002, p. 57). More recently, de Brosses (1765, cf. Ohala & Kawasaki-Fukumori, 1997) presented a three-element hierarchy consisting roughly of stops, liquids and glides, and vowels. He proposed that this ordering produced syllables that were "smooth" (*doux*, in French).

[2] *Sound fullness* (*schallfülle*) in words of Sievers (1876/1893, cf. Parker, 2008).

[3] Several authors consider more complex versions of the hierarchy (Lennertz, 2010; Parker, 2002, 2008), but the one presented here may be considered a minimal sonority hierarchy.

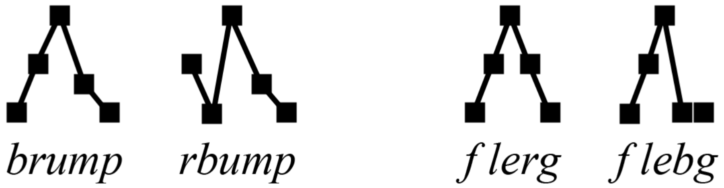

Figure 3.1: Sonority profiles of four monosyllabic words. The Sonority Sequencing Principle states a preference across languages for the profiles of *brump* and *flerg* with respect to the profiles of *rbump*, which has an ill-formed onset, and *flebg*, which has an ill-formed coda.

One of the main constraints that sonority is claimed to impose to all languages is known as the Sonority Sequencing Principle (e.g. see Parker, 2008). This states that in every syllable, there is a unique sonority peak at the syllable nucleus, and then sonority decreases from the nucleus towards the edges of the syllable. For instance, considering syllables with a $C_1C_2VC_3C_4$ structure[4], the Sonority Sequencing Principle states that, *ceteris paribus*, the syllables *brump* and *flerg* are preferred to *rbump* and *flebg*, respectively (see Figure 3.1).

Sonority, as an account of basic phonological constraints common to all languages, has been often challenged (e.g. Harris, 2006; Ohala & Kawasaki-Fukumori, 1997). Criticism come mainly in four forms:

1. Sonority lacks clear acoustic correlates, being instead an *ad hoc* construct.

2. Sonority is a circular explanation because it is used to make inferences about the same set of linguistic facts that were used to propose it.

3. Sonority gives an incomplete account of phonological tendencies that appear to be universal, such as the dispreferred status of clusters like /wu/, /ji/, /tl/.

4. Sonority is inconsistent with a number of languages displaying very often consonant clusters involving sibilants in their syllabic onsets (e.g. *string* [strɪŋ] in English, *sparen*

[4] C stands for any consonant, and V for any vowel.

[ʃpaːrən] in German, and *sbarra* [zbarra] in Italian).

A complete discussion of these arguments is beyond the scope of the current research. However, we need to address briefly these concerns in the context of our work.

Regarding the first point, for a long time no clear acoustic correlates of sonority were agreed upon. However, similar concerns had put in doubt in the past the existence of acoustic correlates of other linguistic concepts such as rhythmic classes of languages (Ramus et al., 1999, proposed a set of acoustic measures that discriminate successfully the "classical" rhythmic classes). This suggests that the quest for acoustic correlates of sonority may just be a matter of time and effort: recently, several authors such as Parker (2008) have proposed acoustic measures that correlate strongly with sonority values.

The circularity of sonority as an explanation of syllabic structure constitutes a concern mostly for research that is purely linguistic. Several psycholinguistic studies have evaluated how sonority affects linguistic performance (e.g. Berent et al., 2008; Ettlinger, Finn, & Hudson Kam, 2011), providing an opportunity to test the predictions based on sonority in a way that is independent of the formulation of the theory. The work presented in this chapter is also framed in this context.

As for the incompleteness of sonority to account for universal phonological restrictions, these criticisms rely on the undemonstrated assumption that all such restrictions stem from a unique mechanism. Although parsimony is a highly desirable characteristic for an account of phonological universal constraints, it is not necessarily the only valid alternative. Sonority accounts typically acknowledge to sibilants (e.g. /s/, /ʃ/, /z/), and particularly to consonant clusters involving sibilants, an exceptional status. However, sonority accounts are not alone in this respect: it has been proposed, for instance, that production of clusters involving sibilants follows a different developmental trajectory than production of other consonant clusters (Fikkert, 1994), and they also receive a special treatment in terms of phonological rules (see

Nespor, 1993, p. 176). Since sonority provides a good account of linguistic preferences as far as sibilants are not concerned, the experimental material used across this chapter will avoid this class of phonemes.

3.1.2 Phonotactics and the Sonority Sequencing Principle

The phonotactic constraints of a language is the set of constraints that a language observes regarding the combination of phonemes. The following are all examples of phonotactic constraints, of different complexities:

- In Standard Italian, no word can finish with a consonant (with a very reduced number of exceptions, such as *con*, *non*, and *per*). Even if Italian dialects are also considered, no word can finish with some consonants like /ɲ/ or /ʃ/ (the initial phonemes of *gnocco* and *sciarpa*, respectively).

- In Spanish, no word can finish with a CC combination (although syllables can: e.g. *cons.ta.tar*).

- In English, words can start with a CCC sequence only if the first consonant is /s/ (e.g. *spring*).

The phonotactics of the native language are acquired during the first years of life, with the earliest experimental demonstrations at about nine months of age (e.g. Archer & Curtin, 2011; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). Before 18 months, infants can learn simple phonotactic regularities from very brief exposure (Chambers, Onishi, & Fisher, 2003), and later they can also learn more complex aspects such as frequency of diphones (Coady & Aslin, 2004).

Given that both of them involve restrictions on phoneme ordering, the distinction between the Sonority Sequencing Principle and phonotactics might at this point get blurry. However, there are crucial aspects that distinguish them: The Sonority Sequencing Principle regards exclusively syllabic structure in terms of the ordering of consonants, and it is believed to be a universal, possibly innate, bias for languages. Instead, phonotactics includes a much wider range of phonological constraints, and it is certainly learned from experience because it is language specific. Nonetheless, sometimes it is not easy to tear apart these two components, such as in the work by MacKenzie, Curtin, and Graham (2012). Among other findings, they observed that 12-month-old English speaking infants failed to associate the Czech words *ptak* and *svet* to novel objects. They interpreted their results as demonstrating that "by 12 months of age, infants are beginning to apply their language-specific phonotactic knowledge to their acceptance of word forms". However, it is possible that this failure is potentiated by (or due to) the ill sonority profile of *ptak*. This problem may be avoided by considering pseudowords with a good sonority profile that violate English phonotactics, such as *tlak*[5].

The relationship between sonority and phonotactics is not rigid, as experience with the native language can reinforce or override the initial biases posed by the sonority hierarchy: If an infant's native language lacks ill-formed syllables (like Spanish or Japanese), then phonotactics will strengthen the bias for well-formed syllables. Alternatively, it is clear that any healthy infant exposed to a language (like Russian or Hebrew) will acquire this language in spite of the several violations to the Sonority Sequencing Principle that it presents.

[5] The cluster /tl/, apart from being a possible syllabic onset according to sonority, is allowed in syllabic onset in a few words of several languages (e.g. *atleta* in Spanish and *genetliaco* in Italian), and in word onset in languages such as Nahuatl, from the Uto-Aztecan family (e.g. *tlacucahuatl*, modified to *cacahuate* in Spanish which means *peanut*).

3.2 Speech perception and the brain

Ever since the pioneering observations by Broca and Wernicke, language has been considered to be lateralized to the left brain hemisphere in the vast majority of human adults, independently of hand preference (Milner, Branch, & Rasmussen, 1964).

It was long known that the anatomy of the human temporal lobe is asymmetric already in newborns (Witelson & Pallie, 1973), and recent *in utero* imaging extended this finding to fetuses (Kasprian et al., 2011). Molfese and Molfese (1979) conducted an early study on the functional specialization of the two brain hemispheres at birth related to language. These authors recorded neonates' auditory evoked potentials during speech perception, observing neural components that differentiated between consonants (/b/ vs. /g/) in the left but not in the right hemisphere. A much stronger proof of functional lateralization for speech perception in human newborns was given by Peña et al. (2003). They used functional near-infrared spectroscopy to reveal that in the first days of life, human infants process differentially speech and non-speech stimuli, with speech involving more strongly perisylvian areas in the left hemisphere and no lateralization for a non-speech control (backward speech). Later studies (e.g. Kotilahti et al., 2010) have corroborated this left-lateralized pattern for speech perception.

Further research suggests that both cerebral hemispheres may be specialized to process auditory stimuli varying in different time scales: Boemio, Fromm, Braun, and Poeppel (2005) observed bilateral responses to auditory stimulation with temporal modulations in the time scale of phonemes, but right-lateralized responses to auditory stimulation whose relevant temporal modulations were slower, in line with the right hemisphere specialization for processing prosodic information (e.g. D. A. Abrams, Nicol, Zecker, & Kraus, 2008; Lindell, 2006). In a similar line of thought, Zatorre and colleagues (e.g. Zatorre, Belin, & Penhune, 2002) had suggested that the the left hemipshere's neural connections make it more apt to process auditory stimuli with high temporal resolution. Telkemeyer et al. (2009) showed that an asymmetric pattern of activity is also present in newborn infants.

Left temporal lobe presents another important ability during the first months of life. Dehaene-Lambertz and Baillet (1998) analyzed 4-month-old infants' mismatch responses in an event-related potential experiment. Infants listened to sequences of syllables $S-S-S-T$, where S stands for a *standard* syllable which is presented three times, and T stands for a *target* syllable which could be either the same as the preceding syllables or a different one. Crucially, the authors used one $S-T$ pair in which both syllables were perceived by French adult listeners as *da*, and another pair in which one member is perceived as *da* and the other as *ba*. Although syllables were designed to match the acoustic change between S and T for both of these pairs, infants' mismatch responses were stronger for the pair that adults perceive as different. The source of this mismatch response was localized on the temporal lobe in the left hemisphere, suggesting that early in life this brain region engages in phonological—as opposed to purely phonetic—processing.

Given all this, we predict that brain activity involving the processing of sonority profiles will particularly engage left temporal areas of the brain.

3.3 Functional near-infrared spectroscopy

In order to study newborns' responses to syllables with different degrees of preference according to the Sonority Hierarchy, we used functional near-infrared spectroscopy (fNIRS).

Functional NIRS is a non-invasive technique for imaging the brain cortex, based on the emission of near-infrared light on the subject's scalp. Near-infrared light has the peculiarity of being able to penetrate human tissues such as skin, bone, and white and grey brain matter. Due to the highly irregular microstructure of some of these tissues, near-infrared light entering the

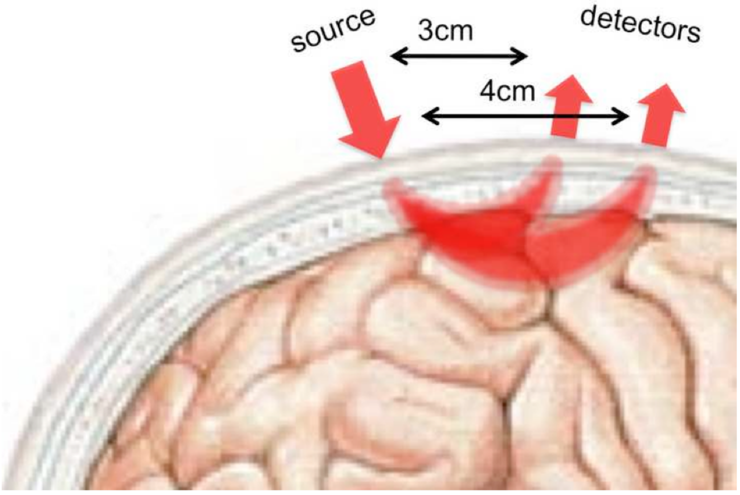

Figure 3.2: How NIRS records hemodynamic activity in the brain cortex. The schema depicts a source, two detectors, and two *channels*, that is the virtual banana-shaped volumes uniting source and detector that are maximally sampled by near-infrared light. Reproduced with permission from Gervain et al. (2011).

human body does not follow a linear trajectory but a curvilinear one. More precisely, near-infrared photons entering in inhomogeneous tissue undergo both absorption and scattering in a different proportion than other photons, meaning that an appreciable part of them will exit from the tissue in a nearby location and very different direction. Typically, near-infrared photons emitted onto the scalp will follow a banana-shaped curve (see Figure 3.2), so that a receptor located just a few centimeters away from the emitter can measure a maximal amount of non-absorbed photons. The virtual volume encompassed by the banana shape connecting emitter and receptor is called a channel.

Measuring the absorption rate of photons in the scalp can give good measurements of the local concentrations of oxy- and deoxy-hemoglobin in the recording channel. It is known that some specific wavelengths are differentially sensitive to the two hemoglobin species, allowing to disentangle their concentration variations in order to estimate the local metabolic consumption evoked by a given task.

In terms of temporal and spatial resolution, fNIRS is located between functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). Several machines for recording fNIRS have a sampling rate in the order of 10 Hz, outperforming state-of-the-art fMRI capabilities[6], and provide a spatial resolution of a few centimeters, more precise than EEG. Moreover, if the intensity of near-infrared light is kept low, fNIRS presents no risk for the subject.

Functional NIRS displays a series of advantages for the study of newborns' brain responses: it makes no noise, permitting the use of auditory stimulation at moderate volume and without the need of earphones. Since most manufacturers embed emitters and receptors of near-infrared light in specialized probes, positioning the optodes on the infant's scalp is a simple and quick process.

The reader is referred to Aslin (2012); Gervain et al. (2011); Lloyd-Fox, Blasi, and Elwell (2010) for more details about how fNIRS works, and for reviews of studies using it for developmental research.

3.4 Experiment 3
Large sonority contrast in the first week of life

We start our exploration of neonatal sensitivity to the Sonority Hierarchy by contrasting two extreme forms of well- and ill-formedness: CCVC syllables with large sonority rises (e.g. *blif*) and large sonority falls (e.g. *lbif*). We additionally look for the brain areas involved in this contrast, assessing specifically the role of left temporal areas of the brain.

[6] This advantage in the time domain must be, however, put into perspective. Notwithstanding the higher sampling rate of fNIRS with respect to fMRI, both rely on the measurement of hemodynamic brain responses. These are slow responses that occur in the time scale of seconds, limiting the potential advantage given by the higher sampling rate.

3.4.1 Participants

Twenty-four healthy newborns (8 boys and 16 girls, aged 2.9 ± 0.83 days) participated in this study. They had Apgar scores of 8.5 ± 0.83 and 9.0 ± 0.20 in the first and fifth minute respectively, a gestational age of 39.3 ± 1.17 weeks, weighted 3.310 ± 0.307 Kg at birth, and had a head circumference of 34.5 ± 1.0 cm.

Additional 8 infants were tested but rejected because of crying or fuzziness ($n = 4$), difficulties in obtaining good signal[7] ($n = 3$), or experimental error ($n = 1$).

3.4.2 Stimuli

All the stimuli used in the experiments of this chapter were naturally recorded. This decision was taken because many of the good speech synthesis tools available today are based on diphone databases, hence presenting a high risk of inserting undesired schwas. This is also a risk with natural recordings, given the difficulty both perceptual and articulatory that ill-formed syllables pose, but this risk can be reduced because there are languages that allow some specific ill-formed syllables, like Russian. Thus, a female native speaker of Russian recorded all the speech material used in Experiments 3, 4, and 5 in this chapter. This material consisted in nonce monosyllabic and bisyllabic words, and it was all recorded in a single session in a sound-attenuated booth. Although not requested specifically to produce infant-directed speech, the speaker had a naturally high voice register[8].

For Experiment 3, we used a set of CCVC syllables with their consonant cluster consisting of either a sonority rise (e.g. “bl”) or a sonority fall (e.g. “lb”). Two sets of 8 words each were

[7] This includes–but is not limited to–infants who have black, thick hair, and infants whose head had a shape or size that did not allow for proper positioning of the NIRS probes.

[8] Would our results have been different if we had used infant-directed speech? It is not clear a priori. Whereas infant-directed speech may improve intelligibility and grab the infant’s attention much more effectively than other types of speech, it seems not to be necessary for successful demonstrations of neonatal capacities. For instance, Gervain, Macagno, Cogoi, Peña, and Mehler (2008) used synthesized speech with flat intonation.

Monosyllables			Bisyllables	
Sonority rises	Sonority plateaus	Sonority falls	Sonority rises	Sonority falls
pras	bkin	rvug	oblif	arbom
fros	dkan	rvem	ivros	ilvan
vlug	gdif	lbug	adrif	urpas
vlin	kdom	rveg	udros	arvud
brud	kvas	rdos	iflud	olbif
vros	kvif	rfug	uprif	olvud
bran	pkan	lbif	afrom	irvug
from	fkom	lvug	oflug	urdos

Table 3.1: Material used in Experiments 3–5.

Wordset	Duration [ms]	Average pitch [Hz]	# of feature changes
Monosyllabic rises	660 (63)	217 (6.3)	3.0 (0.5)
Monosyllabic plateaus	668 (79)	215 (9.9)	3.4 (0.7)
Monosyllabic falls	692 (59)	213 (8.3)	2.9 (1.1)
Bisyllabic rises	778 (39)	229 (8.1)	3.2 (0.9)
Bisyllabic falls	768 (31)	230 (7.9)	3.1 (0.8)

Table 3.2: Basic acoustic and phonetic measurements of the material of Experiments 3–5 (means and standard deviations).

selected (see Table 3.1). Descriptive statistics for duration, average pitch, and number of feature changes between the two consonants in the onset cluster are presented in Table 3.2. The two wordsets did not differ along any of these dimensions (all ps $> .29$). Intensity was set to 70 dB for all words.

3.4.3 Procedure

Newborns were tested in their crib in a silent room located in Udine's Azienda Ospedaliera Santa Maria della Misericordia, either during sleep or in a quiet state of alert. Our procedure is depicted in Figure 3.3, and it follows the block design presented in Benavides-Varela, Gómez, and Mehler (2011). Newborns listened to blocks consisting of either syllables with sonority

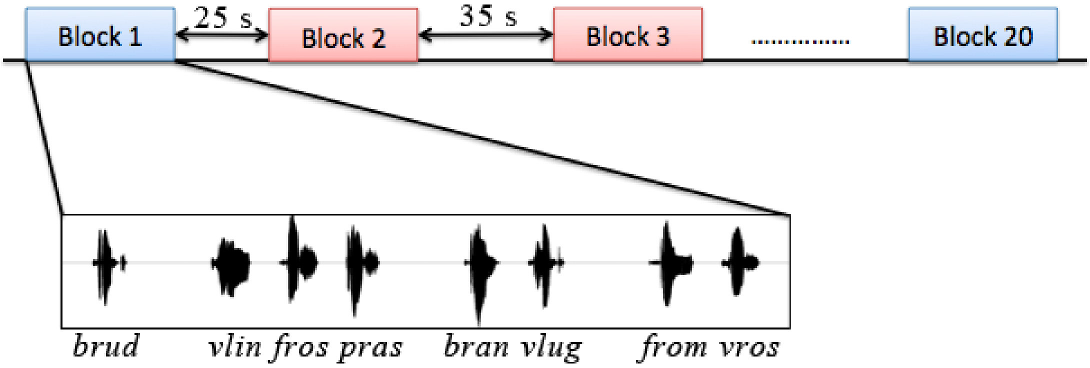

Figure 3.3: Schematic description of the experimental procedure for Experiments 3–5.

rises or syllables with sonority falls. Each stimulation block lasted about 13 s, presenting all eight syllables of a given condition in random order with randomized pauses in between (either 0.5 s or 1.5 s). Blocks were separated by randomized pauses (either 25 s or 35 s), to avoid anticipatory effects previously observed in hemodynamic activity (see for instance Sirotin & Das, 2009, for monkeys, and Nakano, Homae, Watanabe, & Taga, 2008, for 3-month-old infants). We presented ten blocks per condition, for a total duration of about 15 minutes. Order of presentation of blocks was randomized in groups of two, so that in all of these groups there was always one block of each condition.

3.4.4 Data acquisition

Neonates' hemodynamic activity was recorded using fNIRS, by means of a Hitachi ETG-4000 machine (Hitachi Medical Corporation, Tokyo, Japan). This machine emits continuous near-infrared light of two wavelengths (695 and 830 nm) through 10 emitters and records light absorption through 8 detectors. Emitters and detectors are arrayed in two silicon holders (probes) to provide a total of 24 recording channels. Each channel corresponds to an emitter-detector pair having a separation of 3 cm, and sample rate is 10 Hz. Total laser power output

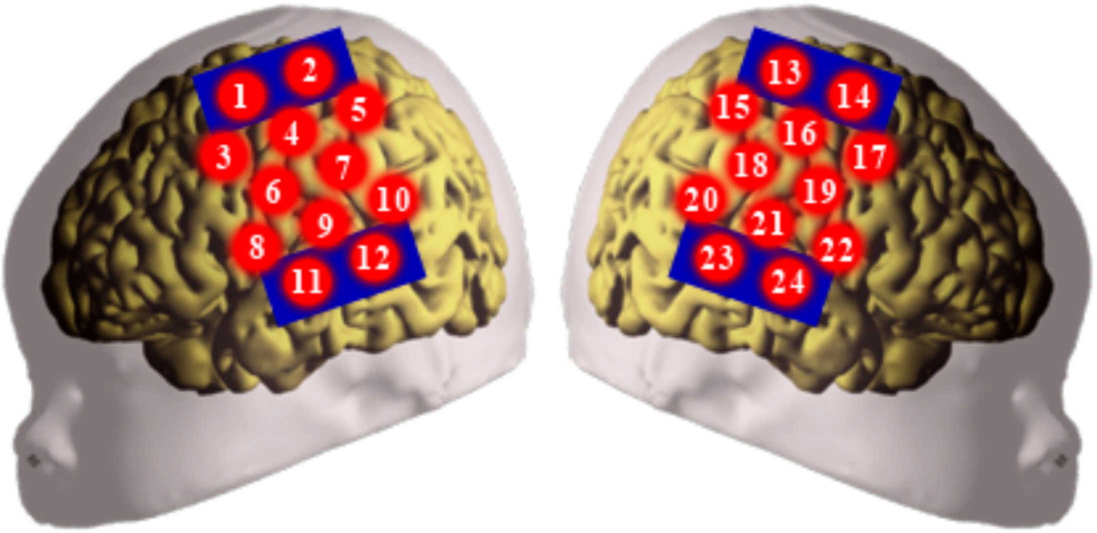

Figure 3.4: Approximate location of the NIRS probes and channels for Experiments 3–5. Blue rectangles show our four regions of interest: left superior (channels 1 and 2), left inferior (channels 11 and 12), right superior (channels 13 and 14), and right inferior (channels 23 and 24).

per optical fiber was 0.75 mW.

The two probes were positioned one over each brain hemisphere by using skull landmarks. The shape and placement of the probes is depicted in Figure 3.4, to maximize the likelihood of recording temporal and perisylvian areas as in Peña et al. (2003).

3.4.5 Data processing and analysis

We analyzed variations in local oxy- and deoxy-hemoglobin concentrations computed from light absorption data collected by the NIRS machine using the modified Beer-Lambert Law (see Gervain et al., 2011). Hemodynamic signals were band-pass filtered between 0.02 Hz and 0.5 Hz to remove slow fluctuations of cerebral blood flow, and other possible artifacts such as heartbeat. Epochs were extracted starting 5 s before each block onset and finishing at 15 s after the end of the block, for a total epoch length of 33 s. For each epoch, signal from specific channels was rejected on the basis of two criteria:

- Saturation of the light measurement, defined as a light measurement higher than the 99% of the maximum recordable. Typically this occurs because a fiber is not contacting properly the neonate's scalp.
- Presence of movement artifacts, defined as changes in the hemodynamic signal larger than 0.1 mmol×mm in a time window of 0.2 s or smaller.

Epochs with more than 12 rejected channels were excluded from analysis. Only participants with at least three good epochs per condition were further considered.

For each good epoch, a linear baseline was fitted and subtracted from the data in order to set to zero the mean activity during the initial and final 5 s portions of the epoch[9]. We analyzed area under the curve for the whole 23 s period in between. Baseline periods were not included in the statistical analysis.

For this experiment and the following ones, we conducted three main analyses:

- A mixed ANOVA with factors Condition (sonority rise vs. sonority fall) and Channel (1, 2, ..., 24), considering subjects as a random factor.
- Paired t-test contrasts of the two conditions on a channel-per-channel basis, correcting for multiple comparisons using the method proposed by Benjamini and Yekutieli (2001, Theorem 1.3). This method controls the False Discovery Rate at a desired α level, by calculating corrected p values from the original ones. Let $p_{(1)}, p_{(2)}, \ldots, p_{(24)}$ be the p values corresponding to the 24 NIRS channels, ordered from smallest to largest. Then,

[9] It has been previously reported that hemodynamic responses reach their maximum and drop back to basal levels in the period between these two portions, so that any non-zero activity out of it can be deemed to be due to baseline fluctuations (e.g. Kotihlahti et al., 2005; Taga & Asakawa, 2007).

for all $k = 1, 2, \ldots, 24$, the k-th corrected p value is computed as

$$\bar{p}_{(k)} = \min\left\{ \frac{24 \cdot c}{k} \cdot p_{(k)} \;\; ; \;\; 1 \right\},$$

where $c = 1 + \frac{1}{2} + \cdots + \frac{1}{24} \approx 3.776$.

- Paired t-test contrasts of the two conditions in four regions of interest (ROIs) corresponding to bilateral temporal and fronto-parietal cortices. Temporal ROIs comprised channels 11 and 12 on the left hemisphere and channels 23 and 24 on the right hemisphere. Fronto-parietal ROIs comprised channels 1 and 2 on the left hemisphere and channels 13 and 14 on the right hemisphere. Focusing on these areas allows us to (a) assess the role of bilateral temporal cortex in the perception of syllables with distinct degrees of well-formedness according to the Sonority Sequencing Principle, evaluating our prediction of a specific engagement of the left hemisphere, and to (b) estimate the contribution of fronto-parietal networks, implicated at birth in several processes related to speech perception such as extraction of regularities (Gervain et al., 2008) and memorization of word forms (Benavides-Varela, Gómez, Macagno, et al., 2011).

  In order to reduce noise in this ROI analysis, we considered only data from infants who had no rejected channels in each given ROI[10]. Given the small number of ROIs, correction for multiple comparisons using False Discovery Rate is less powerful than other methods, so we decided to apply Holm-Bonferroni correction for the ROI analysis. The Holm-Bonferroni method (Holm, 1979) provides strong control of Type-I error, by building corrected p values $\bar{p}_{(1)}, \ldots, \bar{p}_{(4)}$ from the original p values $p_{(1)}, \ldots, p_{(4)}$

[10] Sometimes at the expense of slightly reducing the number of neonates contributing data to each ROI contrast.

(ordered from smallest to largest) in the following way:

$$\bar{p}_{(k)} = \min\left\{ \frac{p_{(k)}}{5-k} \ ; \ 1 \right\}.$$

All analyses were conducted using Matlab R2008b (MathWorks, Inc.).

3.4.6 Results

Figure 3.5 presents the hemodynamic curves for all channels and both conditions. The mixed ANOVA analysis for oxyhemoglobin showed significant main effects of condition ($F(1,1066) = 17.89$, $p < .001$) and channel ($F(23,1066) = 1.81$, $p = .011$), but no interaction ($F(23,1066) < 1$, n.s.). The corresponding analysis for deoxyhemoglobin yielded no significant results (all ps $> .11$). The main effect of condition consisted in higher oxyhemoglobin concentrations for ill-formed syllables.

Paired t-tests on a channel per channel basis revealed significant differences in oxyhemoglobin concentrations between conditions in channels 8 ($p = .047$) and 11 ($p = .024$) on the left hemisphere and channels 13 ($p = .006$) and 14 ($p = .023$) on the right hemisphere, although none of them remained significant after the application of the False Discovery Rate correction (see Section 3.4.5). Again, no differences were detected for deoxyhemoglobin (all uncorrected ps $> .065$).

The ROI analysis yielded more interesting results: oxyhemoglobin responses due to well- and ill-formed syllables differed significantly in Left Inferior ($p = .04$) and Right Superior ($p = .029$) areas (Holm-Bonferroni p values). Deoxyhemoglobin showed no differences for all ROIs (all uncorrected ps $> .33$). Average hemodynamic curves for all ROIs and both hemoglobin species are depicted in Figure 3.6.

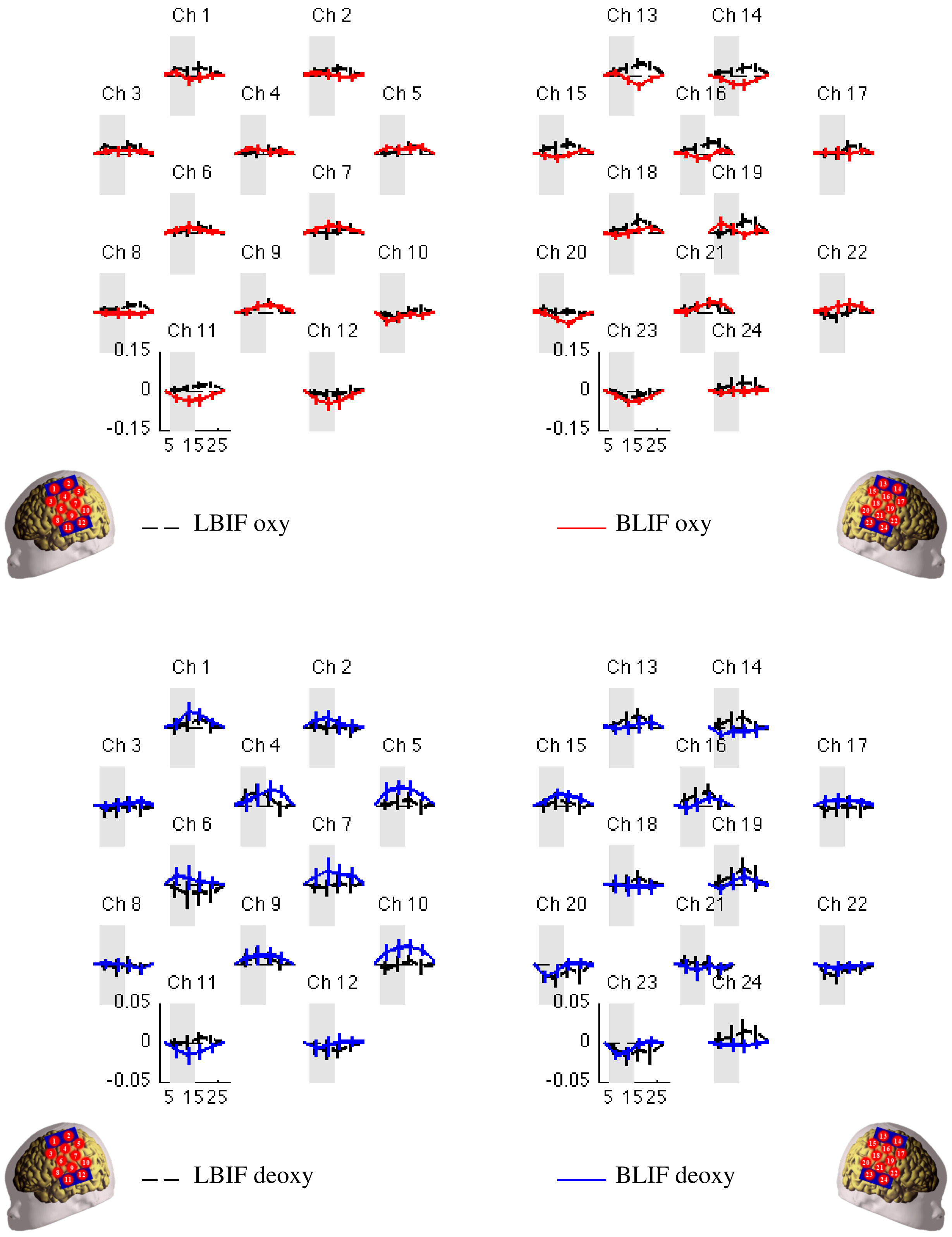

Figure 3.5: Hemodynamic responses for all 24 channels in Experiment 3. Top: Oxyhemoglobin. Bottom: Deoxyhemoglobin. The depicted brains show approximate location of each channel on the newborns' cortex.

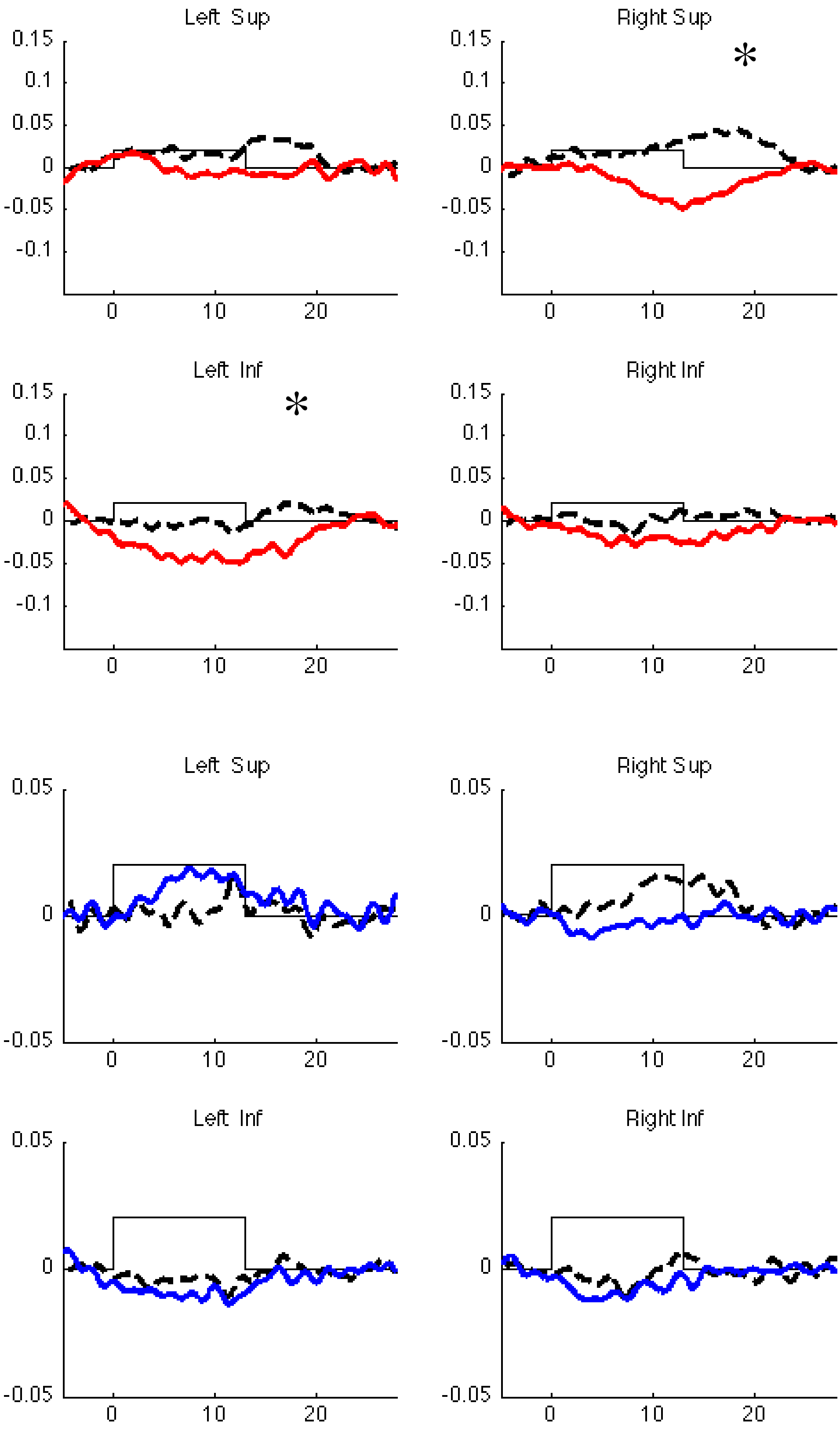

Figure 3.6: Hemodynamic responses for all 4 regions of interest in Experiment 3. Top: Oxyhemoglobin. Bottom: Deoxyhemoglobin. Significant differences in area under the curve are marked by an asterisk.

Region of Interest	Effect size	t-statistic	p value	corrected p value	
Left Superior	$d = 0.29$	$t(18) = 1.25$	$p = .23$	$p = .46$	
Left Inferior	$d = 0.56$	$t(22) = 2.70$	$p = .013$	$p = .04$	*
Right Superior	$d = 0.63$	$t(21) = 2.97$	$p = .007$	$p = .029$	*
Right Inferior	$d = 0.23$	$t(21) = 1.10$	$p = .28$	$p = .28$	

Table 3.3: Statistical analysis for oxyhemoglobin concentrations in Experiment 3 for each region of interest. Significant differences between conditions (after correcting for multiple comparisons) are marked with an asterisk.

3.4.7 Discussion

Experiment 3 evaluated newborn infants' sensitivity to classes of syllables that differ in their degree of preference across languages. Differences in their hemodynamic responses to both syllable types partially support the hypothesis that neonates are sensitive to the sonority profile of classes of syllables. Significant deviations between conditions were found in the Left Inferior and Right Superior regions of interest. Both regions displayed higher concentrations of oxyhemoglobin for ill-formed syllables (strongly dispreferred across languages). We also observed a higher concentration of oxyhemoglobin for ill-formed syllables when considering overall activity (from the mixed ANOVA analysis).

Notice that these hemodynamic responses were elicited by blocks containing eight different words each, hence infants' responses must be based on a structure common to the words in each block, rather than to single items. The two wordsets were controlled for basic acoustic properties: average pitch, duration, intensity, and number of feature changes between the consonants in the onset, discarding these options as drivers of the results. However, it is still unclear whether this common structure extracted by newborns is acoustic or phonological in nature. Moreover, the two aforementioned regions of interest are not necessarily processing the stimuli in the same way.

Differences in the Left Inferior ROI support our hypothesis of involvement of left temporal areas of the brain. Since the left brain hemisphere becomes the main seat of phonological capacities in adulthood, it is at least plausible that this area is responding to phonological structure.

Attested right hemisphere contributions to speech perception tend to be restricted to suprasegmental, intonational properties of language. A work relevant for our purposes is the one by D. A. Abrams et al. (2008): They proved that right-hemisphere auditory cortex is more accurate in representing the speech envelope. Behaviorally, envelope-preserving transformations of speech stimuli allow for better discriminability than other transformations in children and adults (Bertoncini, Serniclaes, & Lorenzi, 2009). A quick look at the envelope of CCVC syllables with sonority rises and falls (see panels A and B of Figure 3.7) suffices to observe that this might be a good alternative explanation to sonority contour: syllables with sonority rises have an envelope with a wide peak (given by the liquid and the vowel, e.g. *li* in *blif*), whereas syllables with sonority falls present a much sharper peak in their envelope (e.g. *i* in *lbif*). Neonates might have succeeded in discriminating well- and ill-formed syllables because of this purely acoustic contrast.

An alternative explanation to Right Superior activity is related to motor theories of speech perception, since the channels composing this ROI are located close to the central sulcus. Works by Pulvermüller et al. (2006) and D'Ausilio et al. (2009) have shown distinct activations in motor cortex associated to the perception of syllables with lip-related and tongue-related consonants, and that disruption of these activations by transcranial magnetic stimulation impedes phoneme perception. In this context, higher metabolic demands for ill-formed syllables might be related to the higher articulatory difficulty of these syllables. Nonetheless, we believe that such interpretation is unlikely because it disregards that motor involvement in speech perception has been established for the left motor cortex, whereas our results point to the right hemisphere. And even if right motor cortex were also involved in speech percep-

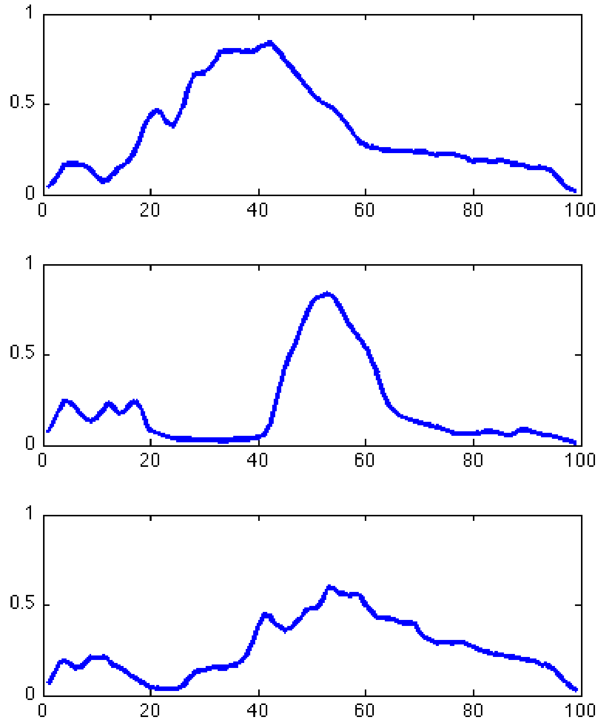

Figure 3.7: Average envelopes for words in the sonority rises, sonority falls, and sonority plateaus sets. Envelope of each word was computed as the amplitude of the complex Hilbert transform of its waveform. Each envelope was then smoothed, normalized in duration, and rescaled to have a maximum value of 1 before averaging.

tion in adulthood (something that very few studies have evaluated), it would be difficult to reconcile with a strong right-hemisphere lateralization at birth.

Yet other account for the involvement of right superior channels is related to the role of right prefrontal cortex in monitoring salient or deviant stimuli (see Vallesi, 2012, and references therein). However, this account has difficulties in at least two grounds: on the one hand, it is improbable that our positioning of the NIRS probes allows for appropriate access to prefrontal cortex. On the other hand, prefrontal cortex matures late in comparison to other brain areas, and adult-like monitoring capacities may not be observable even in 4-year-old children (Vallesi & Shallice, 2007).

The presence of an overall effect of condition indicates that our contrast generated also a degree of non-localized hemodynamic activity. Similar patterns of activity spanning the entire set of NIRS channels have been observed before in studies about neonates' memory for words

(Benavides-Varela, Gómez, Macagno, et al., 2011). There are at least two possible accounts for this phenomenon:

1. This overall hemodynamic pattern reflects hemodynamic activity from the brain. In this case we conclude that a large portion of the brain cortex (more precisely, all brain areas being covered by the 24 NIRS channels of our system) is responding preferentially to one kind of stimulus.

2. This overall hemodynamic pattern reflects hemodynamic activity exterior to the brain. In principle, NIRS measures all hemodynamic changes in the banana-shaped volume connecting emitters and receptors (recall Figure 3.2). This volume includes extra-cranial compartments such as the skin, rich in blood vessels and veins. This would suggest that the sonority contrast is eliciting different responses in the neonates' autonomic nervous system, making the measured response similar in a way to autonomic measures like skin conductance and heart rate[11].

These accounts are not mutually exclusive, but nonetheless both provide further support of the neonates' capacity to discriminate between well- and ill-formed syllables. As a final note, it is hard at this stage to know whether this broad activation is specific to some kind of protocol, equipment, and/or tasks, because most studies do not report main effects of condition spanning the whole set of recording channels (not even some using very similar machinery, protocol, and task, e.g. Gervain et al., 2008; May et al., 2011; Peña et al., 2003).

Concluding, Experiment 3 has showed a sensitivity to at least the acoustic correlates of syllabic well-formedness measured by the sonority hierarchy. This sensitivity is specific to the Left Inferior and Right Superior regions of interest, suggesting the involvement of left temporal and right fronto-parietal cortices.

[11] See Section 3.7.3.

3.5 Experiment 4
A subtler sonority contrast

Experiment 3 indicated both an overall hemodynamic response and a couple of regions in the newborn brain that respond differently to well- and ill-formed syllables. In Experiment 4, we evaluate whether the processes occuring in these regions are better explained on a phonetic or phonological basis. In order to do this, we change our previous condition with sonority falls for one with sonority plateaus. This modification provides a control over onset acoustics (because words from both conditions will have obstruent consonants in their onsets, unlike in Experiment 3) and speech envelope, which will have wider peaks in both conditions (compare panels A and C of Figure 3.7).

3.5.1 Participants

Twenty-four healthy newborns (14 boys and 10 girls, aged 3.0 ± 0.66 days) participated in this study. They had Apgar scores of 8.5 ± 0.51 and 9.3 ± 0.44 in the first and fifth minute respectively, a gestational age of 38.9 ± 1.50 weeks, weighted 3.429 ± 0.378 Kg at birth, and had a head circumference of 35.0 ± 0.8 cm.

Additional 16 infants were tested but rejected because of crying or fuzziness ($n = 10$) or difficulties in obtaining good signal ($n = 6$).

3.5.2 Stimuli

The same speaker who recorded the stimuli of Experiment 3 also recorded a set of CCVC syllables featuring sonority plateaus in their consonant cluster (e.g. *bd*). From there, a set of 8 words was selected for presentation in this experiment (see Table 3.1). The set of syllables with sonority rises was the same as in Experiment 3. Descriptive statistics for duration, average pitch, and number of feature changes between the two consonants in the onset cluster for both conditions are presented in Table 3.2. The two wordsets did not differ in any of these dimensions (all ps $> .21$). Intensity was set to 70 dB for all words.

3.5.3 Procedure

The procedure was identical to the one used in Experiment 3, presenting syllables with sonority plateaus instead of syllables with sonority falls.

3.5.4 Data acquisition

Data acquisition was identical to Experiment 3.

3.5.5 Data processing and analysis

Data processing and analysis were identical to Experiment 3.

3.5.6 Results

Figure 3.8 presents the hemodynamic curves for all channels and both conditions. The ANOVA analysis for oxyhemoglobin concentrations yielded a main effect of condition ($F(1,1039) = 5.64$, $p = .018$), a trend to significance for the main effect of channel ($F(23,1039) = 1.52$, $p = .056$), and no interaction ($F(23,1039) = 1.01$, $p = .45$). The corresponding analysis for deoxyhemoglobin revealed a main effect of channel ($F(23,1039) = 1.66$, $p = .027$), with neither a main effect of condition ($F(1,1039) = 2.32$, $p = .13$) nor interaction ($F(23,1039) = 1.19$, $p = .24$).

Region of Interest	Effect size	t-statistic	p value	corrected p value	
Left Superior	$d = 0.51$	$t(19) = 2.27$	$p = .035$	$p = .11$	
Left Inferior	$d = 0.60$	$t(22) = 2.88$	$p = .009$	$p = .035$	*
Right Superior	$d = 0.44$	$t(17) = 1.86$	$p = .081$	$p = .16$	
Right Inferior	$d = 0.21$	$t(19) = 0.96$	$p = .35$	$p = .35$	

Table 3.4: Statistical analysis for oxyhemoglobin concentrations in Experiment 4 for each region of interest. Significant differences between conditions (after correcting for multiple comparisons) are marked with an asterisk.

Paired t-tests revealed significant differences only in channel 12 (p = .004) on the left hemisphere for oxyhemoglobin, and in channel 18 (p = .046) on the right hemisphere for deoxyhemoglobin. As in Experiment 3, these results were not significant when correcting for multiple comparisons.

The ROI analysis for oxyhemoglobin yielded significant differences for both Left Inferior and Superior areas (see Table 3.4), of which only the Left Inferior remained significant after Holm-Bonferroni correction[12]. As depicted in Figure 3.9, this difference is due to syllables with sonority plateaus eliciting higher hemodynamic responses than than syllables with sonority rises. The corresponding analysis for deoxyhemoglobin revealed no significant differences (all uncorrected ps $> .38$).

3.5.7 Discussion

In Experiment 4 we observed no difference between conditions in the Right Superior region of interest. On the contrary, we observed distinct hemodynamic activity for well- and ill-formed syllables in the Left Inferior region, replicating the results of Experiment 3. Additionally, the overall difference in hemodynamic activity when considering all 24 recording sites was also

[12] As it can be seen in Figure 3.9, sometimes the superior regions of interest appeared to have large differences. These differences were not always significant as it was the case for the Right Superior area (both for oxy- and deoxy-hemoglobin). The reason behind this was a high variability of the hemodynamic responses between subjects in those regions.

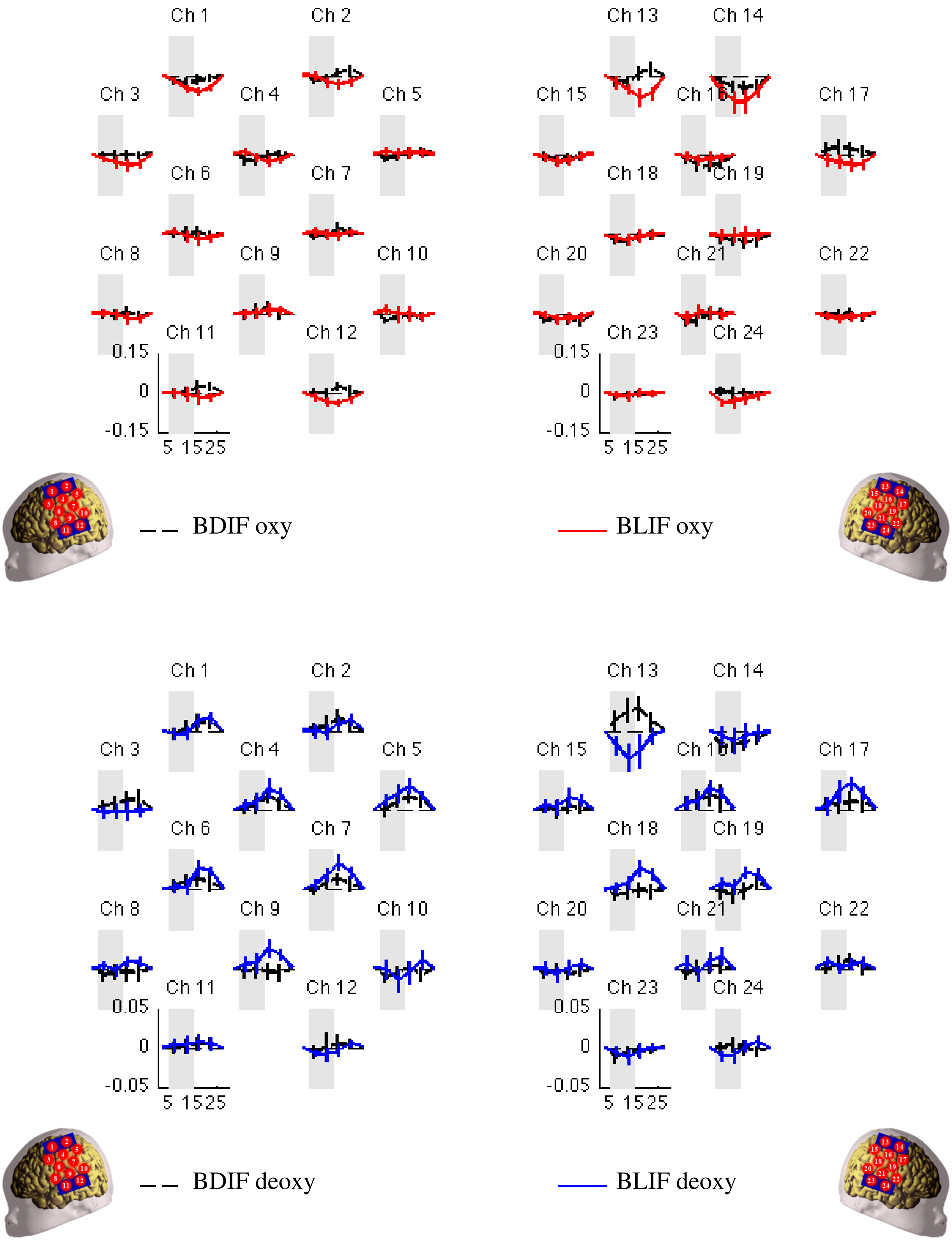

Figure 3.8: Hemodynamic responses for all 24 channels in Experiment 4. Top: Oxyhemoglobin. Bottom: Deoxyhemoglobin. The depicted brains show approximate location of each channel on the newborns' cortex.

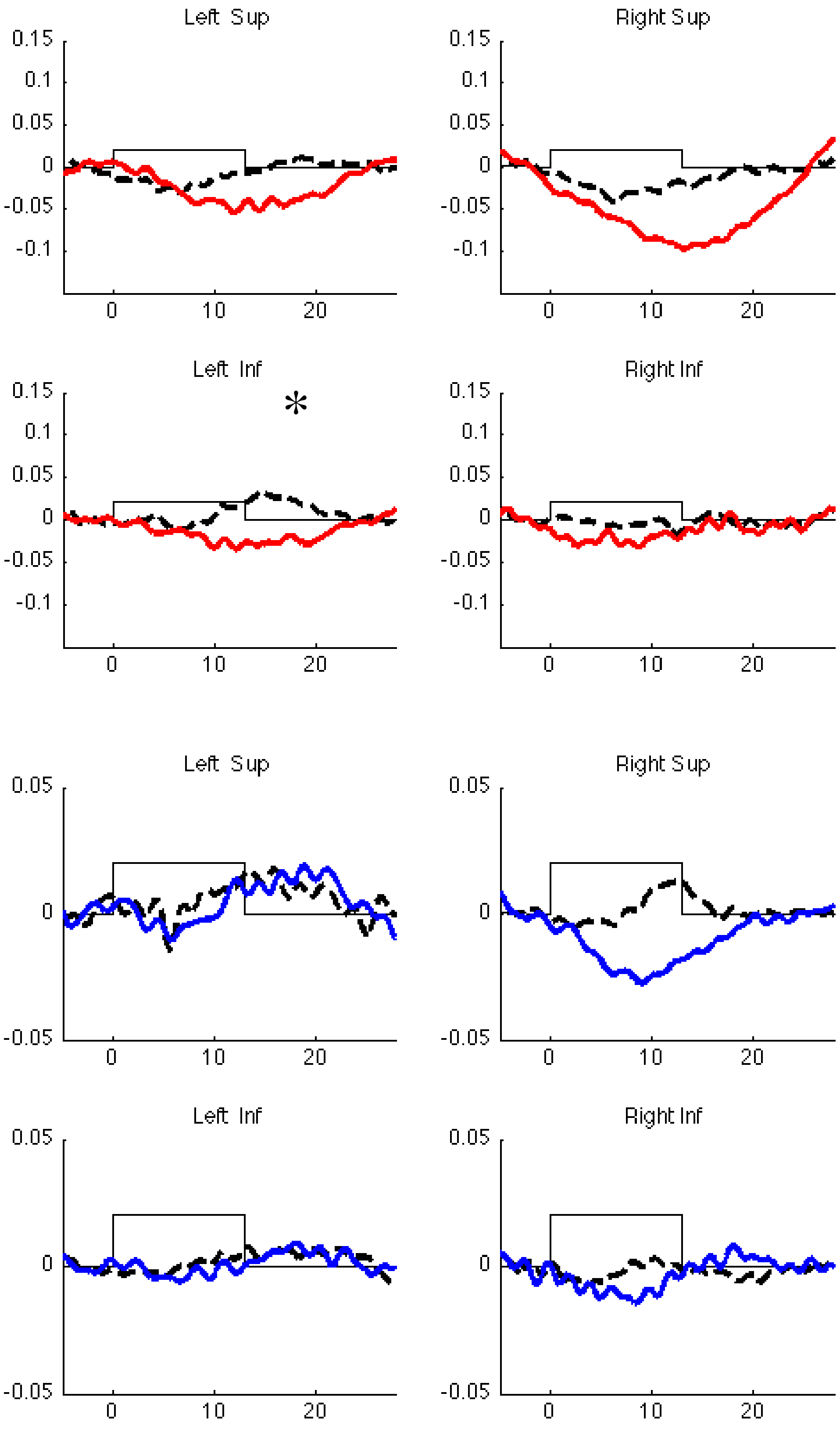

Figure 3.9: Hemodynamic responses for all 4 regions of interest in Experiment 4. Top: Oxyhemoglobin. Bottom: Deoxyhemoglobin. Significant differences in area under the curve are marked by an asterisk.

replicated.

The absence of effect in Right Superior areas is not explainable simply in terms of statistical power: although this region of interest showed a reduced number of degrees of freedom with respect to Experiment 3 (17 against 21, due to a higher rejection rate of channels in that region), there was also a reduction of the effect size of the order of 30%. This suggests that the change in hemodynamic activity in that region between experiments truly relates to the change in the experimental conditions. Whereas the well-formed syllables were the same as in our previous experiment, Experiment 4 included a set of CCVC syllables with sonority plateaus. These syllables present waveforms better matched to well-formed syllables in acoustic terms (onset consonants and envelope), indicating that activity in Right Superior areas in Experiment 3 related mainly to the acoustic contrasts between syllable types instead of other accounts like monitoring[13].

The effect previously found in the Left Inferior region was replicated, with an almost equal effect size (variation of a 7%). Moreover, the effect had the same direction, with dispreferred syllables eliciting higher concentrations of oxyhemoglobin with respect to preferred syllables. All together, this supports the hypothesis that this region is processing speech stimuli on a primarily phonological basis and, moreover, responding to well-formedness in terms of sonority.

Experiment 4 also replicated the overall hemodynamic pattern contrasting our two conditions, underlining neonates' capacity of discriminating speech sounds on the basis of their sonority profile. For a discussion on the possible interpretations of this effect, see Section 3.7.3.

[13] Notice that even if one of the channels included in the Right Superior region of interest apparently displays differences between conditions in Experiment 4 (see Figure 3.8), this channel (number 13) is the one located more posteriorly. This channel covers more likely fronto-parietal regions instead of frontal and prefrontal regions, suggesting that executive functions are not the driving force behind the Right Superior activation in Experiment 3.

Still, it is possible that these results do not refer to the organization of speech units larger than pairs of phonemes. Infants might respond with lower concentrations of oxyhemoglobin to syllables with sonority rises because these appear more frequently than sonority falls or plateaus in their environment. Almost all of our participants were born to Italian-speaking families[14], and Italian lacks ill-formed syllables. Although Italian allows sonority falls and plateaus (e.g. *albero*, *espurgare*, and *subdolo*) when considering any pair of adjacent phonemes, these are never contained in a single syllable. Up to now, we have no evidence that the hemodynamic responses in Experiments 3 and 4 are not due to the pairs of phonemes themselves, instead of the pairs and their specific location within the word. We will address this concern in our last experiment of this chapter.

3.6 Experiment 5
Large sonority contrast in a bisyllabic context

Experiment 4 suggested that neonates' responses to monosyllabic words with different degrees of well-formedness according to the Sonority Hierarchy are composed of two components: one occurring in left temporal cortex, probably phonological in nature, and one occurring across the whole set of recording sites.

Experiment 5 will explore whether the results obtained so far stem from the organization of speech units in the newborn brain or, alternatively, they reflect statistics on phonemic pairs learned from exposure to Italian. Specifically, Experiment 5 will compare words with consonants clusters with sonority rises and sonority falls as in Experiment 3, but word structure will be now VCCVC, —that is, bisyllabic—and the cluster will no longer be at word onset.

[14] Precisely, all but one family in Experiment 3, one in Experiment 4, and two in Experiment 5.

That is, we will now contrast words like *oblif* and *olbif* in the sonority rise and fall conditions, respectively. The distribution of phoneme pairs in these word classes is comparable–though not identical–to the one of Experiment 3, and thus a phoneme-pair-based approach predicts that the results of Experiment 5 should be similar to those of Experiment 3. On the other hand, an approach based on the Sonority Sequencing Principle predicts that infants will parse the bisyllabic words so as to avoid the presence of ill-formed units, rendering words of both conditions well formed. This implies that Experiment 5 should yield null results.

3.6.1 Participants

Twenty-four healthy newborns (10 boys and 14 girls, aged 3.2 ± 0.80 days) participated in this study. They had Apgar scores of 8.6 ± 0.7 and 9.1 ± 0.5 in the first and fifth minute respectively, a gestational age of 39.3 ± 1.3 weeks, weighted 3.359 ± 0.398 Kg at birth, and had a head circumference of 34.7 ± 1.1 cm.

Additional 18 infants were tested but rejected because of crying or fuzziness ($n = 11$) or difficulties in obtaining good signal ($n = 7$).

3.6.2 Stimuli

The same speaker who recorded the stimuli of Experiments 3 and 4 also recorded a set of VCCVC bisyllabic words featuring sonority rises or sonority falls in their consonant cluster (e.g. *oblif* and *olbif*, respectively). From there, two sets of 8 words were selected for presentation in this experiment (see Table 3.1). Descriptive statistics for duration, average pitch, and number of feature changes between the two consonants in the onset cluster for both conditions are presented in Table 3.2. The two wordsets did not differ in any of these dimensions (all ps $> .58$). Intensity was set to 70 dB for all words.

3.6.3 Procedure

The procedure was identical to the one used in Experiment 3, presenting VCCVC bisyllabic words with sonority rises and falls instead of monosyllables.

3.6.4 Data acquisition

Data acquisition was identical to Experiment 3.

3.6.5 Data processing and analysis

Data processing and analysis were identical to Experiment 3.

3.6.6 Results

The ANOVA analysis for oxyhemoglobin revealed a main effect of condition ($F(1,1073) = 8.18$, $p = .004$), and no effect of channel ($F(23,1073) = 1.33$, $p = .14$) or interaction ($F(23,1073) < 1$, n.s.). This time the main effect of condition was reversed: words with sonority rises were associated to higher oxyhemoglobin concentrations than words with sonority falls. The corresponding analysis for deoxyhemoglobin yielded no significant results (all ps $> .31$).

Paired t-tests for all channels yielded differences in channels 9 ($p = .033$) in the left hemisphere and 23 ($p = .043$) in the right hemisphere for oxyhemoglobin, and in channel 1 ($p = .029$) in the left hemisphere for deoxyhemoglobin. None of these results remained significant after controlling for multiple comparisons.

The ROI analysis revealed no significant differences in oxyhemoglobin concentrations (all uncorrected ps $> .29$, see Table 3.5). Instead, deoxyhemoglobin concentrations differed

Region of Interest	Effect size	t-statistic	p value
Left Superior	$d = -0.10$	$t(21) = -0.46$	$p = .65$
Left Inferior	$d = 0.16$	$t(21) = 0.77$	$p = .45$
Right Superior	$d = -0.25$	$t(19) = -1.10$	$p = .29$
Right Inferior	$d = -0.23$	$t(21) = -1.09$	$p = .29$

Table 3.5: Statistical analysis for oxyhemoglobin concentrations in Experiment 5 for each region of interest.

Region of Interest	Effect size	t-statistic	p value	corrected p value
Left Superior	$d = -0.45$	$t(21) = -2.12$	$p = .046$	$p = .18$
Left Inferior	$d = 0.27$	$t(21) = 1.26$	$p = .22$	$p = .66$
Right Superior	$d = -0.19$	$t(19) = -0.86$	$p = .40$	$p = .80$
Right Inferior	$d = 0.16$	$t(21) = 0.75$	$p = .46$	$p = 1$

Table 3.6: Statistical analysis for deoxyhemoglobin concentrations in Experiment 5 for each region of interest.

in the Left Superior region ($t(21) = -2.12$, $p = .046$), although it was not significant after Holm-Bonferroni correction.

3.6.7 Discussion

Experiment 5 evaluated the possibility that brain responses obtained in Experiment 3 were due to statistics based on phonemic pairs inherent to syllables containing consonantal clusters with sonority rises and sonority falls. To do this, we added a vowel at the beginning of monosyllables similar to those of Experiment 3. This manipulation resulted in phonemic pair sets comparable to those of Experiment 3, but presenting bisyllabic words in both conditions. Crucially, from the point of view of the Sonority Sequencing Principle, now the two conditions are composed of well-formed words.

Results of Experiment 5 turned null in every region of interest. Although we are not statistically entitled to take this as support for the sonority hypothesis, the clear contrast between the results of Experiments 3 and 5 is suggestive.

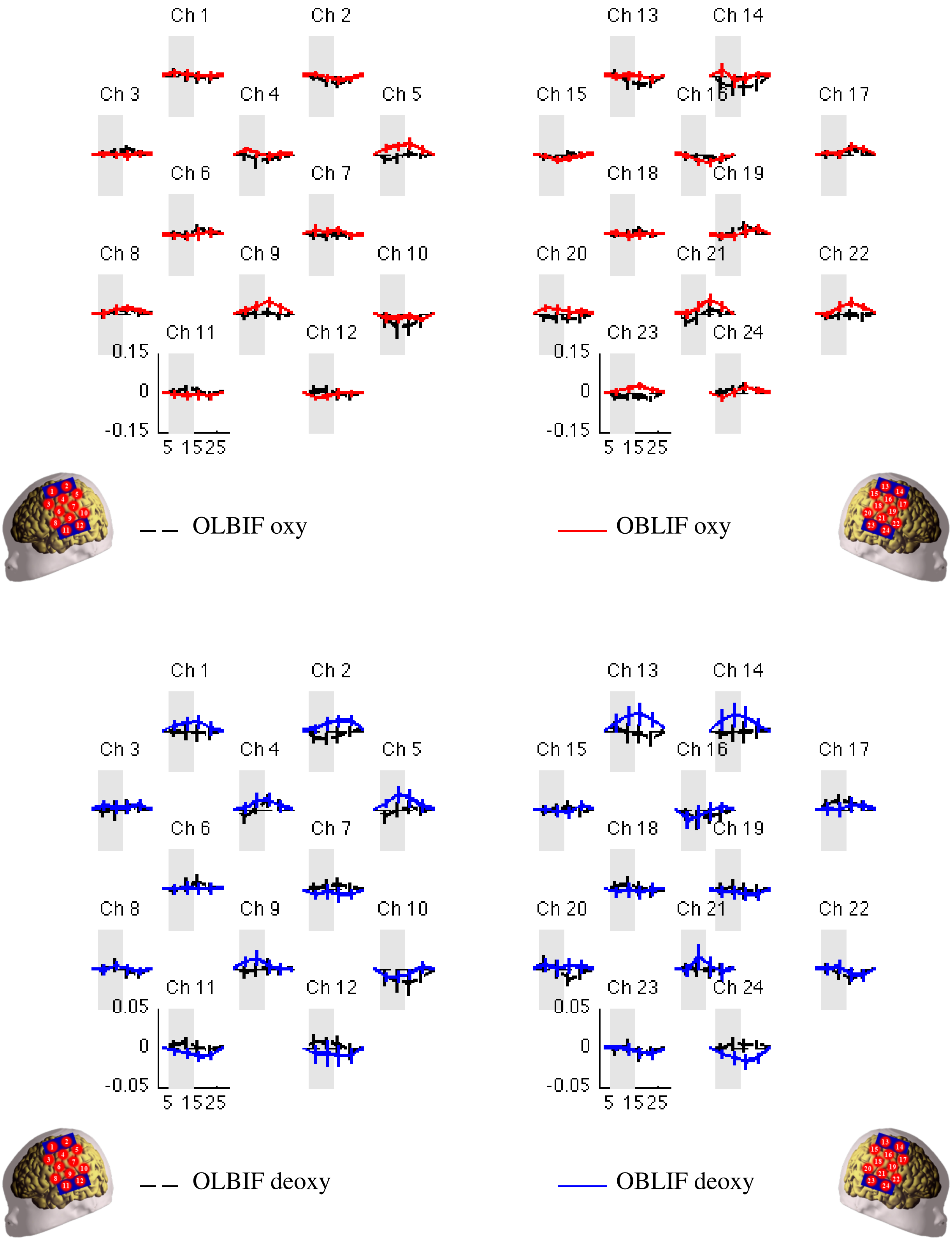

Figure 3.10: Hemodynamic responses for all 24 channels in Experiment 5. Top: Oxyhemoglobin. Bottom: Deoxyhemoglobin. The depicted brains show approximate location of each channel on the newborns' cortex.

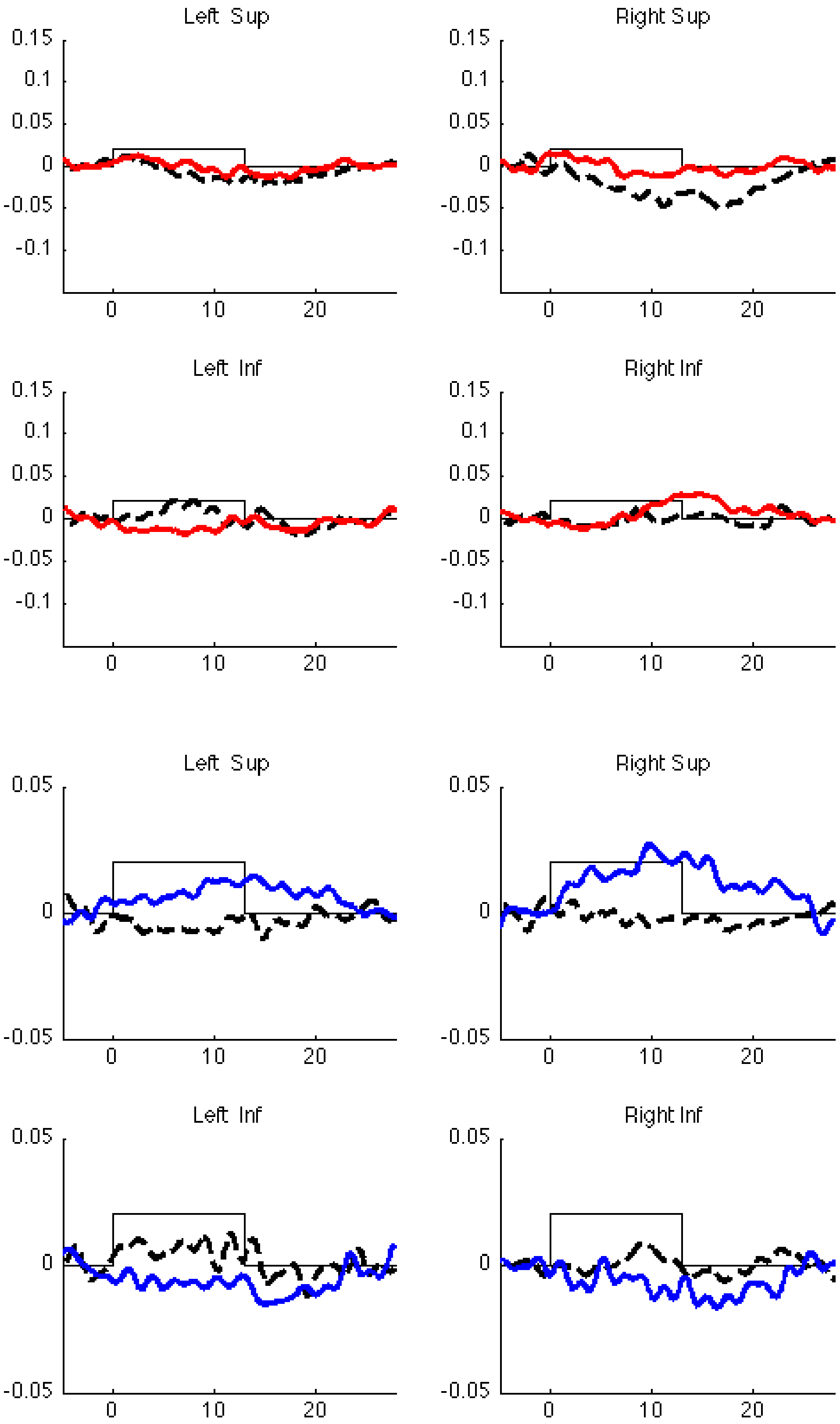

Figure 3.11: Hemodynamic responses for all 4 regions of interest in Experiment 5. Top: Oxyhemoglobin. Bottom: Deoxyhemoglobin.

Somewhat surprisingly, not all relevant contrasts were null in Experiment 5: we still observed an effect of condition in the overall level of oxyhemoglobin, but of reversed sign with respect to Experiment 3. That is, words containing a sonority rise (e.g. *oblif*) elicited higher concentrations of oxyhemoglobin than words containing a sonority fall (e.g. *olbif*). This reversal cannot be explained either in terms of adjacent phonemes or in terms of the Sonority Sequencing Principle, because the latter concerns only the ordering of consonants within a single syllable. However, this reversal further underlines that the difference between the two conditions observed in Experiment 3 has changed in Experiment 5, as predicted by sonority.

3.7 Chapter discussion

In this chapter, we have presented three experiments suggesting that human newborns are sensitive to the Sonority Sequencing Principle just as human adults are, by:

1. Demonstrating that newborns display different hemodynamic responses to classes of syllables sharing their sonority contour.

2. Observing that, consistently, syllables with ill sonority profiles elicit higher concentrations of oxyhemoglobin in channels located over left temporal cortex.

3. Showing that these responses are not due to statistics based on adjacent phonemes, and that they interact with the syllabic structure of the material.

Experiment 3 showed that newborns distinguish well- and ill-formed syllables (e.g. *blif* vs. *lbif*). We stress the fact that neonates were presented with a set of eight different words in each condition, so that these results reflect generalization of properties common to all syllables in each condition, instead of the specific acoustic properties of single items. The sets of words for both conditions were matched for syllable duration, average pitch, average intensity, and number of consonantal feature changes in the consonant cluster, discarding all of these low-level cues as drivers of our results. Significant differences across conditions were observed in two regions of interest formed by left inferior and right superior channels, in addition to a broad hemodynamic pattern of higher oxyhemoglobin for ill-formed syllables.

Still, well- and ill-formed syllables differ in many aspects, two of the most salient being their temporal envelope and their onset acoustics. This issue was addressed in Experiment 4, where the contrast of sonority rises and plateaus (e.g. *blif* vs. *bdif*) controls for both dimensions. Results of Experiment 4 confirmed the hemodynamic pattern observed in left inferior channels in Experiment 3 with a different contrast and a new set of subjects, lending credibility to the interpretation of left temporal activity as reflecting syllabic well-formedness instead of just acoustics. Moreover, the broad hemodynamic pattern was also replicated.

Finally, Experiment 5 explored whether the results of Experiment 3 occur specifically when the consonant cluster occurs in syllable-initial position, or also in word-medial position (e.g. *oblif* vs. *olbif*). Crucially, the addition of a vowel at the beginning of an ill-formed syllable allows for resyllabification of the consonant cluster, converting the ill-formed syllable in a well-formed bisyllable. Neonates' hemodynamic activity did not show differences across regions of interest, although a broad hemodynamic pattern differed, in the opposite direction with respect to the previous experiments.

In the next sections, we discuss specific topics regarding our results and methodology.

3.7.1 Sonority or novelty

The presence of higher hemodynamic responses for ill-formed syllables in Experiments 3 and 4 could be interpreted as a novelty effect instead of a reflection of an early constraint in the perception of speech. Such explanation entails that infants have partially learned the relative frequency of consonantal pairs before the end of their first week of life. Given that the vast majority of our young participants were born to Italian-speaking families, and Italian shows a strong bias against ill-formed syllables, it is pertinent to ask whether learning that well-formed syllables are preferred to other kinds of syllables might have been achieved by this time.

These are two non-exclusive alternatives for such learning to occur. Preferences for well-formed syllables might have been learned prenatally, or during the first hours of life.

Several works indicate that fetuses learn intonational and rhythmic characteristics of their native language and react according to them after birth (Mampe et al., 2009; May et al., 2011; Moon, Cooper, & Fifer, 1993). However, fetuses would require much more than access to suprasegmental properties of Italian in order to display a bias for well-formed syllables: They would need to perceive consonantal segments—or at least classes of consonantal segments such as obstruents and liquids—. Research on the acoustic properties of the womb (e.g. in humans, Querleu, Renard, Versyp, Paris-Delrue, & Crèpin, 1988; in sheep, R. M. Abrams et al., 1998) cast doubt on this possibility, since intrauterine sound levels are increasingly attenuated for frequencies above 300 Hz, making consonant discrimination particularly difficult[15].

The degraded consonantal information that the fetus receives makes it unlikely that prenatal speech perception is the source of our results. The other possibility, as we mentioned earlier, is that infants learned the sonority constraints out of the womb during the first hours or days of life. Nonetheless, we believe this interpretation is unlikely to be correct: If that were the case, our findings would amount to the earliest evidence so far of phonotactic learning, pushing the bar down from 9 months to 5 days of age.

It is relevant to notice that the level of phonotactic complexity required to learn sonority

[15] Shannon, Zeng, Kamath, Wygonski, and Ekelid (1995) showed that spectral information from three broad frequency bands may be enough for adult subjects to recognize consonants presented in an /aCa/ context. Still, their findings support the hypothesis that spectral information above 300 Hz is required, since even the lowest frequency band considered by them spanned up to 800 Hz.

constraints is higher than simply prohibiting clusters like *lb*. A correct formulation should forbid clusters like *lb* specifically in syllable-initial position (or at least, word-initially). Moreover, since we presented 8 different words per condition in all our experiments, neonates would have needed to learn the relevant information for many consonantal pairs or to have generalized the dispreferred status of, say, liquid-obstruent clusters in initial position with respect to the reverse. Although theoretically possible, we believe parsimony advises against this learning explanation.

3.7.2 Pooled analysis of the well-formedness contrast

By pooling together the data from Experiments 3 and 4, we can conduct a more powerful analysis of the sonority contrast, regardless of the degree of ill-formedness. For this, we looked for significant clusters of channels whose hemodynamic activity induced by well- and ill-formed syllables differ. We used the clustering method presented in page 48.

Based on the spatial disposition of emitters and receptors in each silicon probe, we assume two channels to be neighbors if they share either emitter or receptor. This gives the connectivity graph depicted in Figure 3.12.

For our NIRS data, we used a threshold of $\tau = 2.3$, revealing a significant cluster in the left hemisphere showing differences between conditions, composed by channels 11 and 12 ($p = .046$). This provides support both to our selection of the regions of interest analyzed in this chapter, and to our results that suggest lateralization of the sonority contrast.

3.7.3 Possible involvement of the autonomic nervous system

It becomes necessary at this point to give a critical and careful look to the possibility, outlined in the discussion of Experiment 3, that the overall hemodynamic responses differing across

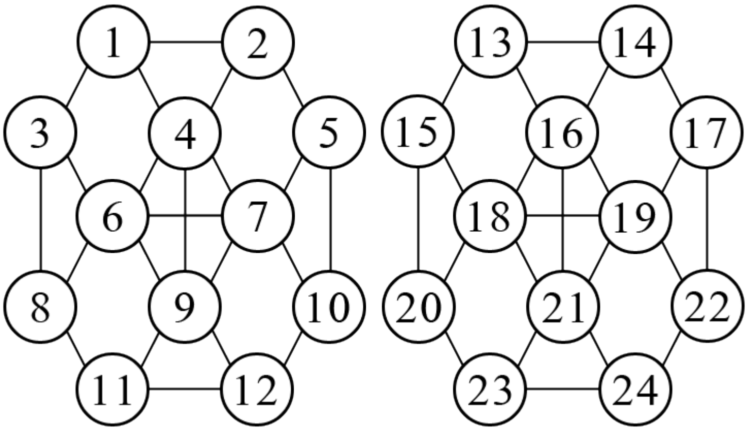

Figure 3.12: Connectivity graph used for the cluster analysis. Two channels were considered as connected if they shared either emitter or receptor.

conditions reflect activity of the autonomic nervous system.

First of all, we notice that the plausibility of observing widespread cortical activation in response to our material is not very high, since perception of spoken words is expected to involve more specific areas of the brain cortex as attested by many studies using NIRS. Additionally, clinical and cognitive researchers have studied for decades fetal and neonatal responses to auditory stimuli in terms of systemic measures such as heart rate (e.g. Clarkson & Berg, 1983; DeCasper, Lecanuet, Busnel, Granier-Deferre, & Maugeais, 1994; Ockleford, Vince, Layton, & Reader, 1988).

Our NIRS probes were located on the sides of the head, from right above the ears and extending to the vertex. The scalp in this area is irrigated by the superficial temporal artery and the posterior auricular artery. By means of these arteries, the autonomic nervous system could induce measurable changes in blood oxygenation via sympathetic vasoconstriction, sympathetic vasodilatation, or parasympathetic vasodilatation (Franchini & Cowley Jr., 2004), which could be associated to Sokolov's defensive (vasoconstriction) and orienting (vasodilatation) reflexes (see Graham & Jackson, 1970, for a short review). The pattern

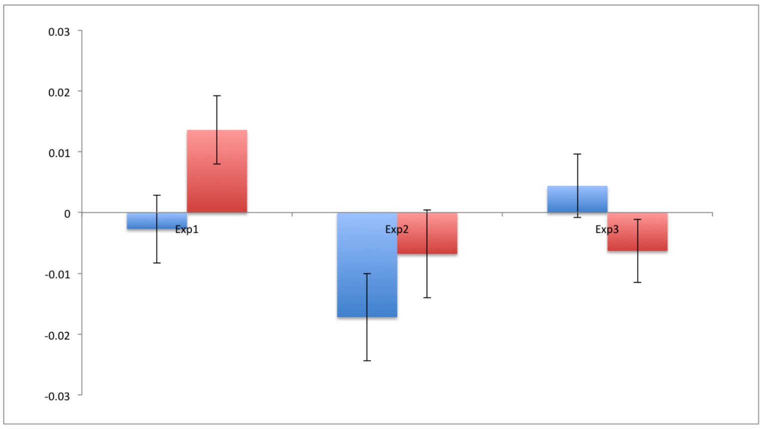

Figure 3.13: Average oxyhemoglobin levels across Experiments 3–5, reflecting the main effects of condition evidenced by the ANOVA analysis. Blue columns represent conditions *blif* (Exp. 3), *blif* (Exp. 4), and *oblif* (Exp. 5). Red columns represent conditions *lbif* (Exp. 3), *bdif* (Exp. 4), and *olbif* (Exp. 5). Vertical lines depict standard errors.

of results across experiments, however, indicates that the responses to monosyllables with sonority rises are less stable than the differences across conditions. Figure 3.13 shows that responses to monosyllables containing sonority rises were close to zero (Experiment 3) or negative (Experiment 4), whereas monosyllables with sonority falls elicited positive changes and monosyllables with sonority plateaus are in between. This suggests the involvement of both vasoconstriction and vasodilatation mechanisms.

Nonetheless, NIRS was not primarily intended to record autonomic responses, which makes all our conclusions tentative. Future research should evaluate autonomic responses by means of heart rate measurements or laser Doppler flowmetry in co-registration with NIRS, in order to clarify the differential involvement of localized and systemic responses.

3.7.4 A wider sonority framework

Independently of the interpretation of the main effect of condition found in Experiments 3–5, the question about the reversal of this effect remains. In Experiments 3 and 4, the main effect of condition revealed higher oxyhemoglobin concentrations for sonority falls and plateaus, whereas in Experiment 5 this main effect showed higher concentrations for sonority rises.

A tentative answer for this question may come from a more general sonority framework. As we mentioned in the introduction to this chapter, the Sonority Sequencing Principle is the most relevant and well-known set of constraints based on the Sonority Hierarchy. There is, however, another constraint in this group called the Syllable Contact Law (see Parker, 2008, p. 56), which states that if a consonant cluster occurs at the juncture between two syllables, then a sonority fall is preferred to a sonority rise. The Syllable Contact Law is a logical counterpart of the Sonority Sequencing Principle: the former states that sonority falls are preferred in syllabic junctions, whereas the latter asserts that sonority rises are preferred in syllabic onsets.

This explanation seems promising because it potentially gives an account of the pattern reversal in the results of Experiment 5 with respect to the other experiments. However, we point out that this is valid only under some assumptions on newborns' perception of our bisyllabic material. Words like *olbif* could in principle be perceptually parsed by newborns as either *o.lbif*, *ol.bif*, or *olb.if*, and similarly words like *oblif* could be parsed as either *o.blif*, *ob.lif*, or *obl.if*. The Sonority Sequencing Principle precludes *o.lbif* and *obl.if* as possible parsings, but still leaves ambiguities. If newborns were shown to perceive *ol.bif* and *ob.lif*, then the Syllable Contact Law would readily imply a preference for the former, unifying the results of Experiments 3–5 under the account "oxyhemoglobin concentration are lower for preferred items according to sonority principles". This is beyond what we can state from our results, and further research should evaluate the validity of sonority predictions on the perception of syllabic contact[16].

[16] We are unaware of any such investigation, although proposals in this direction have already been done (e.g. Peperkamp, 2007).

What our results do suggest is that the neural mechanisms underlying this putative sensitivity to the Syllable Contact Law must be different from the ones of the Sonority Sequencing Principle, because Experiment 5 showed no trace of relevant activity occurring in left inferior channels (in opposition to Experiments 3 and 4). A conjecture about this difference in activity patterns may be related to the very nature of the material involved. Experiments 3 and 4 presented monosyllabic words whose relevant differences were interior to a syllable. This is the domain of the Sonority Sequencing Principle, and Experiments 3 and 4 consistently showed both localized and broad results. On the contrary, the relevant contrast of Experiment 5 involved integration of elements belonging to two different syllables (their edge phonemes).

3.7.5 Cortical involvement

We want to comment on another important issue regarding our results: several of our recorded responses do not resemble a canonical BOLD response, according to well-established fMRI and fNIRS adult findings. Figure 3.14 depicts a typical adult hemodynamic response measured by NIRS: a transient increase of oxy-hemoglobin accompanied by a decrease of deoxy-hemoglobin.

In infant research, results regarding the shape of the hemodynamic response have been mixed: sometimes the measured increase in oxy-hemoglobin is not accompanied by a corresponding decrease in deoxy-hemoglobin (e.g. Gervain et al., 2008), or stimulation induces local decreases–rather than increases–in oxy-hemoglobin (e.g. Watanabe, Homae, & Taga, 2011).

Recent research (Gagnon et al., 2012; Kirilina et al., 2012; Minati, Kress, Visani, Medford, & Critchley, 2011) has suggested that NIRS may be highly sensitive to hemodynamic activity not originating in the brain but in the skin. The work by Kirilina et al. (2012) is

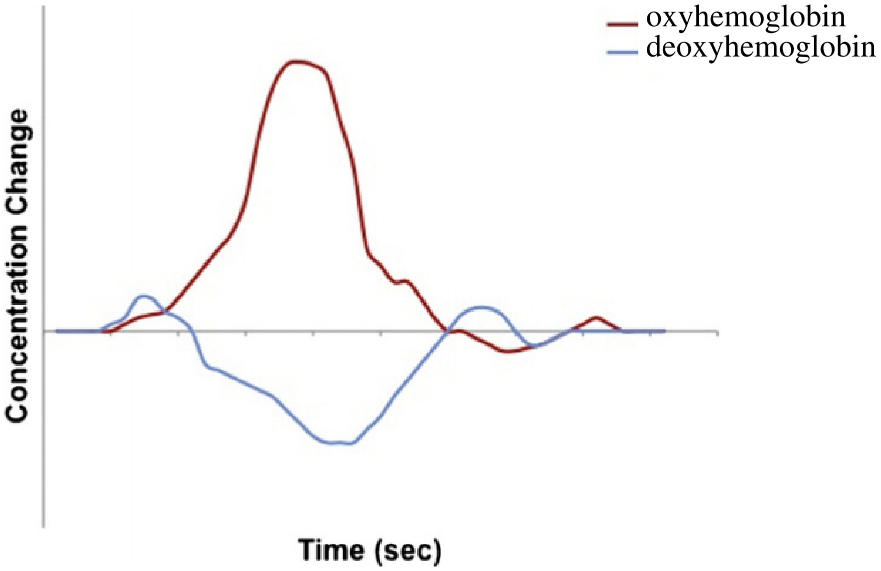

Figure 3.14: Schematic representation of the shape of an ideal hemodynamic response. Reproduced with permission from Gervain et al. (2011).

particularly enlightening in this respect: They recorded fNIRS and fMRI from adults' prefrontal cortex, and observed that fNIRS signals resembled more fMRI signals localized in extracranial space than fMRI signals localized in the brain cortex.

Still, if the different responses we observed between well- and ill-formed syllables on the left inferior region were only due to autonomic responses, it would be hard to explain why these responses are asymmetric. Our sample sizes were not small with respect to the observed effect sizes, thus type I errors are not a very plausible explanation. Moreover, the cluster analysis with the pooled data presented in Section 3.7.2 was consistent in suggesting left lateralization.

3.7.6 Final remarks

We have shown that newborns' perception of speech is modulated by universal phonological preferences predicted by the sonority hierarchy, suggesting that the primordial mechanisms of segmentation of speech are subject to inborn structural constraints. An exploration of the neural bases of these constraints points to left temporal cortex as the seat of sonority-based discriminations, and to the presence of widespread effects possibly due to systemic responses to well- and ill-formed speech items.

Our results raise the question on how these early constraints on speech perception might shape the evolution of spoken languages.

Still, we acknowledge that the distinction between phonology and acoustics in early infancy represents a challenging enterprise. Infants perceive and discriminate a wealth of acoustic contrasts that all adults who have learned a language have lost, but at the same time they display evidence of phonologically-based generalization (such as the abstraction of phonemic categories across speakers, Dehaene-Lambertz & Peña, 2001). It is important that future research tackles this issue and tests more comprehensively the contributions of phonetics and phonology to the generalizations based on sonority profile that we have uncovered.

Chapter 4

General Discussion

4.1 Summary of results

Every newborn infant faces the challenge of selecting, segmenting, and parsing the most relevant signals embedded in their surrounds. Infants particularly excel at tuning their perceptual systems in order to grasp the multi-level regularities of languages, spoken or signed.

This thesis explored two different levels of the hierarchical phenomenon that language is.

- Chapter 2 addressed the process of segmenting words from a continuous stream of syllables. Working with adult participants, we evaluated two possible online measures of the segmentation process, one behavioral and one neural.

  The behavioral measure consisted in observing the reaction times to detect an extraneous signal in the speech stream. This method was inspired by a tradition of psycholinguistic research that proposed that monitoring techniques could reveal the organization into units of the speech stream. We inserted clicks both between and within the words of an artificial language, and observed that participants took longer to detect clicks internal to word units. This pattern emerged smoothly in time, suggesting that segmentation is an incremental process as a well-known model based on computation of distributional information (Aslin et al., 1998; Saffran, Newport, & Aslin, 1996) proposes.

To obtain a neural measure, we looked at brain oscillations during a segmentation task. Our main result pointed to activity in theta band as a relevant process. The patterning of theta oscillations suggested a dual-stage underlying process, with a fast initial stage dependent on attention, and a longer subsequent stage of enhancement of memory traces. This process reminds another model present in the literature (Perruchet & Vinter, 1998), in which chunks of syllables are first stored in a proto-lexical space to be then strengthened or weakened depending on their recurrence in the input.

- Chapter 3 explored the organization of the primordial syllable, that is the syllabic perception that the naïve infant brain deploys before significant exposure to segmental properties of language and before learning the articulatory restrictions imposed by a vocal tract.

  We asked whether newborn infants' perception of isolated syllables would be affected by organizing phonological principles that apply across spoken languages of the world. Universal linguistic principles represent an opportunity to approach the primordial state of the language faculty before significant exposure to language takes place. The specific principle we evaluated is called Sonority Sequencing, which states that syllables are segmented from the speech stream so as to contain a single sonority peak (a vowel, most of the times) flanked by phonemes with decreasing sonority values. When syllables are presented in isolation, this principle implies that syllables not displaying this sonority profile are dispreferred.

  Three experiments used functional near-infrared spectroscopy to probe hemodynamic responses of the newborn brain to classes of syllables with different sonority profiles.

The first two experiments showed a consistent hemodynamic pattern where syllables violating the Sonority Sequencing Principle elicited higher concentrations of oxyhemoglobin when compared to syllables with a good sonority profile. Moreover, this difference was observed in channels located over left temporal cortex, an area of the brain that has been shown critical for speech processing. The last experiment indicated that this hemodynamic pattern does not replicate when one of the offending segments is allowed to move to a neighboring syllable, suggesting that the newborn brain reacts specifically to sequences of segments that unambiguously constitute syllables with an ill-formed sonority profile.

We additionally observed in these experiments a main effect of condition irrespective of the recording site. This second hemodynamic pattern hints at a possible systemic response led by the autonomic nervous system. Strikingly, the putative systemic response was also responsive to the sonority profiles presented in each experiment.

4.2 Implications for language acquisition

Infants have already made considerable progress in acquiring the language(s) of their surrounds by the time they utter their first words. In particular, they have tuned their perceptual systems to process the speech signal at multiple levels. Our results shed light on the inner workings of segmentation processes: from syllables to words, and from phonemes to syllables.

We have explored the time course of word segmentation from continuous speech, shedding light on the online evolution of this process. Our neural data suggests the existence of two stages in its unfolding, something that has been frequently overlooked because dominant offline methods are blind to the process but look only at its outcomes. These two stages are reminiscent of encoding and consolidation processes in memory, and their further study will surely improve our understanding of word segmentation.

We have also demonstrated that some segmentation processes are guided by organizing constraints such as the Sonority Sequencing Principle from the first days of life. Our findings suggest that the quantity and quality of exposure to speech required for the emergence of sonority-related restrictions is much smaller than previously thought. Newborn infants count only on two main sources of linguistic experience by the first week of life: prenatal exposure to their maternal language (poor in terms of the quality of segments, because of the filtering properties of the womb) and exposure to full-fledged speech in their first days of life (poor in terms of quantity). Moreover, our findings rule out the hypothesis that experience with language production is necessary at all.

All together, this thesis work has explored aspects of the multi-level segmentation process. We have demonstrated that some constraints are available to infants from their very first experiences with language. We also explored the mechanisms of word segmentation and observed a dual-stage process, in which each stage likely has distinct attentional and computational requirements.

4.3 Future lines of work

All scientific enterprises leave unanswered questions. In this section, we propose a few directions for future work that we deem relevant for the understanding of language at different levels of its processing. Some of these have already been proposed in the respective chapters, and we summarize them here.

4.3.1 On the assembly of syllables into words

- Although a good first approximation, real continuous speech does not come in 3-min-long excerpts lacking pauses. An event-related design based on the presentation of short sentences of an artificial language would be a much more natural context where to test segmentation. Even if such designs raise several confounds[1], we believe that they can inform successfully about segmentation in a ecological context. If applied to an EEG experiment, such design would also be an improvement in terms of signal-to-noise ratio.

- It is also important to replicate and understand better our results on the time course of spectral power. If our interpretation is correct, theta power should exhibit a two-stage behavior even with more complex artificial languages. Additionally, we predict a differential requirement of attentional resources for the two stages: high attentional demands for the initial, encoding phase, and low for the subsequent, strengthening phase.

4.3.2 On the assembly of phonemes into syllables

- The overall hemodynamic response that we observed in Experiments 3–5 is suggestive, but not conclusive, regarding an involvement of the autonomic nervous system in modulating the peripheral blood flow. This hypothesis can be tested directly, by measuring peripheral blood flow with laser Doppler flowmetry.

- Some languages do present violations of the Sonority Sequencing Principle, not only allowing sonority plateaus but even sonority falls in word onsets, as well. How do in-

[1] Splitting continuous speech into sentences would introduce new segmentation cues such as the appearance of sentence edges. Listening to the artificial sentence *pulikipelubamusomita*, for instance, makes certain that *pu* is a possible word onset. Moreover, this kind of information has a very different time course to distributional cues: edge cues are immediate, whereas distributional cues must be extracted from repeating instances. This makes difficult to evaluate the contribution of the slower segmentation mechanisms. The use of long continuous speech streams has successfully avoided this problem, at the expense of ecological validity.

fants learn that these structures are acceptable? Exposure to such syllabic forms should start early, but there are no works yet estimating the frequency with which infants (and children) hear them.

4.3.3 On the simultaneous execution of both processes

- Ettlinger et al. (2011) asked adult participants to segment a continuous speech stream with some syllables having complex onsets. These syllables had different sonority profiles in their onsets, ranging from large sonority rises (e.g. dnɛku) to large sonority falls (e.g. lbɪzo). They observed that, when exposed to a continuous version of the speech stream, English listeners' performance followed closely the predictions of the Sonority Sequencing Principle: words with sonority rises were successfully segmented, whereas words with sonority falls were below chance.

  Ettlinger et al. asked participants to assembly units at two levels simultaneously: from phonemes to syllables, and from syllables to words. Their results underline the functional relevance of the Sonority Sequencing Principle for adults presented with consonant clusters unattested in their native language but, as with previous research on sonority, the role of experience with language production is not addressed. Our results with neonates suggest that Ettlinger et al.'s findings should hold in preverbal infants, as well.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. Chicago: Aldine.

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience*, *20*(6), 952–964.

Abla, D., & Okanoya, K. (2009). Visual statistical learning of shape sequences: An ERP study. *Neuroscience Research*, *64*(2), 185–190.

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *The Journal of Neuroscience*, *28*(15), 3958–3965.

Abrams, R. M., Griffiths, S. K., Huang, X., Sain, J., Langford, G., & Gerhardt, K. J. (1998). Fetal music perception: The role of sound transmission. *Music Perception*, *15*(3), 307–317.

Archer, S. L., & Curtin, S. (2011). Perceiving onset clusters in infancy. *Infant Behavior and Development*, *34*(4), 534–540.

Aslin, R. N. (2012). Questioning the questions that have been asked about the infant brain using near-infrared spectroscopy. *Cognitive Neuropsychology*, *29*(1-2), 7–33.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

Bastiaansen, M. C. M., van der Linden, M., ter Keurs, M., Dijkstra, T., & Hagoort, P. (2005).

Theta responses are involved in lexical-semantic retrieval during language processing. *Journal of Cognitive Neuroscience*, *17*(3), 530–541.

Benavides-Varela, S., Gómez, D. M., Macagno, F., Bion, R. A. H., Peretz, I., & Mehler, J. (2011). Memory in the neonate brain. *PLoS ONE*, *6*(11), e27497. (doi: 10.1371/journal.pone.0027497)

Benavides-Varela, S., Gómez, D. M., & Mehler, J. (2011). Studying neonates' language and memory capacities with functional near-infrared spectroscopy. *Frontiers in Psychology (Language Sciences)*, *2:64*. (doi: 10.3389/fpsyg.2011.00064)

Benjamini, Y., & Yekutieli, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

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

Berent, I. (2009). Unveiling phonological universals: A linguist who asks "why" is (inter alia) an experimental psychologist. *Behavioral & Brain Sciences*, *32*(5), 450–451.

Berent, I., Lennertz, T., Jun, J., Moreno, M. A., & Smolensky, P. (2008). Language universals in human brains. *Proceeding of the National Academy of Sciences of the USA*, *105*(14), 5321–5325.

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, *104*(3), 591–630.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceeding of the National Academy of Sciences of the USA*, *109*(9), 3253–3258.

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: rhythmical basis of speech representations in neonates. *Language and Speech*, *38*(4), 311–329.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, *4*, 247–260.

Bertoncini, J., Serniclaes, W., & Lorenzi, C. (2009). Discrimination of speech sounds based upon temporal envelope versus fine structure cues in 5- to 7-year-old children. *Journal of Speech, Language, and Hearing Research*, *52*(3), 682–695.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*(4), 711–721.

Bion, R. A. H., Benavides-Varela, S., & Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Language and Speech*, *54*(1), 123–140.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, *8*, 389–395.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(6), 451–459.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2007). On consonants, vowels, chickens, and eggs. *Psychological Science*, *18*(10), 924–925.

Bond, Z. S. (1972). Phonological units in sentence perception. *Phonetica*, *25*, 129–139.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Buiatti, M., Peña, M., & Dehaene-Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage*, *44*(2), 509–519.

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1971). *Problems of knowledge and freedom*. New York: Pantheon Books.

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon Books.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1–22.

Clarkson, M. G., & Berg, W. K. (1983). Cardiac orienting and vowel discrimination in newborns: Crucial stimulus parameters. *Child Development*, *54*(1), 162–171.

Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, *89*(3), 183–213.

Cohen, L., & Mehler, J. (1996). Click monitoring revisited: An on-line study of sentence comprehension. *Memory & Cognition*, *24*(1), 94–102.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorial in Quantitative Methods for Psychology*, *1*(1), 42–45.

Cowie, F. (2010). Innateness and language. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2010 ed.). `http://plato.stanford.edu/archives/sum2010/entries/innateness-language/`.

Crain, S., & Pietroski, P. (2001). Nature, nurture and Universal Grammar. *Linguistics and Philosophy*, *24*(2), 139–186.

Cunillera, T., Càmara, E., Toro, J. M., Marco-Pallares, J., Sebastián-Gallés, N., Ortiz, H., et al. (2009). Time course and functional neuroanatomy of speech segmentation in adults. *NeuroImage*, *48*(3), 541–553.

Cunillera, T., Toro, J. M., Sebastián-Gallés, N., & Rodríguez-Fornells, A. (2006). The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Research*, *1123*(1), 168–178.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, *96*(3), 233–262.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the english vocabulary. *Computer Speech and Language*, *2*(3-4), 133–142.

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, *20*, 1–10.

Cutler, A., Kearns, R., Norris, D., & Scott, D. R. (1993). Problems with click detection: Insights from cross-linguistic comparisons. *Speech Communication*, *13*, 401–410.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*(4), 385–400.

Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing* (pp. 113–134). Hillsdale, NJ: Lawrence Erlbaum.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*(5), 381–385.

de Brosses, C. (1765). *Traité de la formation méchanique des langues, et de principes physiques de l'étymologie*. Paris: Chez Saillant, Vincent, Desaint.

De Diego Balaguer, R., Toro, J. M., Rodríguez-Fornells, A., & Bachoud-Lévi, A.-C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS ONE*, *2*(11), e1175. (doi: 10.1371/journal.pone.0001175)

DeCasper, A. J., Lecanuet, J.-P., Busnel, M.-C., Granier-Deferre, C., & Maugeais, R. (1994). Fetal reactions to recurrent maternal speech. *Infant Behavior and Development*, *17*(2), 159–164.

Dehaene-Lambertz, G., & Baillet, S. (1998). A phonological representation in the infant brain. *NeuroReport*, *9*(8), 1885–1888.

Dehaene-Lambertz, G., & Peña, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *NeuroReport*, *12*(14), 3155–3158.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*, 79–82.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van Der Vreken, O. (1996). The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the fourth International Conference on Spoken Language Processing* (Vol. 3, pp. 1393–1396). Philadelphia.

Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, *105*(3), 1901–1911.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*, *134*(3), 406–419.

Ettlinger, M., Finn, A. S., & Hudson Kam, C. L. (2011). The effect of sonority on word segmentation: Evidence for the use of a phonological universal. *Cognitive Science*, *36*(4), 655–673.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral & Brain Sciences*, *32*, 429–492.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language. *Current Anthropology*, *46*(4), 621–646.

Fell, J., Fernández, G., Klaver, P., Elger, C. E., & Fries, P. (2003). Is synchronized neuronal gamma activity relevant for selective attention? *Brain Research Reviews*, *42*(3), 265–272.

Fikkert, P. (1994). *On the acquisition of prosodic structure*. The Hague: Holland Academic Graphics.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458–467.

Fitch, W. T. (2011). Unity and diversity in human language. *Philosophical Transactions of the Royal Society B (Biological Sciences)*, *366*, 376–388.

Fitzgibbon, S. P., Pope, K. J., Mackenzie, L., Clark, C. R., & Willoughby, J. O. (2004). Cognitive tasks augment gamma EEG power. *Clinical Neurophysiology*, *115*(8), 1802–1809.

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, *4*, 414–420.

Fodor, J. A., Bever, T. G., & Garrett, M. F. (1974). *The psychology of language*. New York: McGraw-Hill.

Foss, D. J., & Lynch, R. H. J. (1969). Decision processes during sentence comprehension: Effects of surface structure on decision times. *Perception & Psychophysics*, *5*, 145–148.

Franchini, K. G., & Cowley Jr., A. W. (2004). Neurogenic control of blood vessels. In D. Robertson (Ed.), *A primer on the autonomic nervous system* (pp. 139–143). Boston, MA: Academic Press.

Franco, A., Cleeremans, A., & Destrebecqz, A. (2011). Statistical learning of two artificial languages presented succesively: How conscious? *Frontiers in Psychology (Language Sciences)*, *2*, 229. (doi: 10.3389/fpsyg.2011.00229)

Frauenfelder, U., Segui, J., & Mehler, J. (1980). Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior*, *19*, 328–337.

French, R. M., & Perruchet, P. (2009). Generating constrained randomized sequences: Item frequency matters. *Behavior Research Methods*, *41*(4), 1233–1241.

Gagnon, L., Yücel, M. A., Dehaes, M., Cooper, R. J., Perdue, K. L., Selb, J., et al. (2012). Quantification of the cortical contribution to the NIRS signal over the motor cortex using concurrent NIRS-fMRI measurements. *NeuroImage*, *59*(4), 3933–3940.

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty.* `http://www.ling.upenn.edu/~ycharles/papers/quick.pdf`. (Manuscript, Yale University)

Garrett, M. F., Bever, T. G., & Fodor, J. A. (1966). The active use of grammar in speech perception. *Perception & Psychophysics*, *1*, 30–32.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceeding of the National Academy of Sciences of the USA*, *105*(37), 14222–14227.

Gervain, J., Mehler, J., Werker, J. F., Nelson, C. A., Csibra, G., Lloyd-Fox, S., et al. (2011). Near-infrared spectroscopy: A report from the McDonnell infant methodology consortium. *Developmental Cognitive Neuroscience*, *1*(1), 22–46.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, *20*(2), 229–252.

Grabner, R. H., Brunner, C., Leeb, R., Neuper, C., & Pfurtscheller, G. (2007). Event-related eeg theta and alpha band oscillatory responses during language translation. *Brain Research Bulletin*, *72*(1), 57–65.

Graham, F. K., & Jackson, J. C. (1970). Arousal systems and infant heart rate responses. *Advances in Child Development and Behavior*, *5*, 59–117.

Gómez, C. M., Vaquero, E., López-Mendoza, D., González-Rosa, J., & Vásquez-Marrufo, M. (2004). Reduction of EEG power during expectancy periods in humans. *Acta Neurobiologiae Experimentalis*, *64*(2), 143–151.

Hald, L. A., Bastiaansen, M. C. M., & Hagoort, P. (2006). Eeg theta and gamma responses to semantic violations in online sentence processing. *Brain & Language*, *96*(1), 90–105.

Hanslmayr, S., Staudigl, T., & Fellner, M.-C. (2012). Oscillatory power decreases and long-term memory: The information via desynchronization hypothesis. *Frontiers in Human Neuroscience*, *6*, 74. (doi: 10.3389/fnhum.2012.00074)

Harris, J. (2006). The phonology of being understood: Further arguments against sonority. *Lingua*, *116*(10), 1483–1494.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and vowels: Different roles in early language acquisition. *Developmental Science*, *14*(6), 1445–1458.

Hochmann, J.-R., Endress, A. D., & Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition*, *115*(3), 444–457.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Holmes, V. M., & Forster, K. I. (1970). Detection of extraneous signals during sentence recognition. *Perception & Psychophysics*, *7*, 297–301.

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(6), 1143–1154.

Huang, R.-S., Jung, T.-P., & Makeig, S. (2009). Tonic changes in eeg power spectra during simulated driving. In D. Schmorrow, I. Estabrooke, & M. Grootjen (Eds.), *Foundations of augmented cognition. neuroergonomics and operational neuroscience* (Vol. 5638, pp. 394–403). Springer Berlin / Heidelberg.

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, *130*(4), 658–680.

Ihara, A., Hirata, M., Sakihara, K., Izumi, H., Takahashi, Y., Kono, K., et al. (2003). Gamma-band desynchronization in language areas reflects syntactic process of words. *Neuroscience Letters*, *339*(2), 135–138.

Jensen, O., & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *European Journal of Neuroscience*, *15*(8), 1395–1399.

Johnson, E. K., & Tyler, M. D. (2009). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23.

Jusczyk, P. W., Cutler, A., & Redanz, N. (1993). Preference for the predominant stress patterns of english words. *Child Development*, *64*(3), 675–687.

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by infants. *Developmental Psychology*, *23*(5), 648–654.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*(3), 402–420.

Jusczyk, P. W., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive Psychology*, *39*(3/4), 159–207.

Kaiser, J., & Lutzenberger, W. (2005). Cortical oscillatory activity and the dynamics of auditory memory processing. *Reviews in the Neurosciences*, *16*(3), 239–254.

Kasprian, G., Langs, G., Brugger, P. C., Bittner, M., Weber, M., Arantes, M., et al. (2011). The prenatal origin of hemispheric asymmetry: An in utero neuroimaging study. *Cerebral Cortex*, *21*(5), 1076–1083.

Kirilina, E., Jelzow, A., Heine, A., Niessing, M., Wabnitz, H., Brühl, R., et al. (2012). The physiological origin of task-evoked systemic artefacts in functional near infrared spectroscopy. *NeuroImage*, *61*(1), 70–81.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.

Kirmizi-Alsan, E., Bayraktaroglu, Z., Gurvit, H., Keskin, Y. H., Emre, M., & Demiralp, T. (2006). Comparative analysis of event-related potentials during Go/NoGo and CPT: Decomposition of electrophysiological markers of response inhibition and sustained attention. *Brain Research*, *1104*(1), 114–128.

Klimesch, W., Doppelmayr, M., Russegger, H., & Pachinger, T. (1996). Theta band power in the human scalp eeg and the encoding of new information. *NeuroReport*, *7*(7), 1235–

1240.

Kotihlahti, K., Nissilä, I., Huotilainen, M., Mäkelä, R., Gavrielides, N., Noponen, T., et al. (2005). Bilateral hemodynamic responses to auditory stimulation in newborn infants. *NeuroReport*, *16*(12), 1373–1377.

Kotilahti, K., Nissilä, I., Näsi, T., Lipiäinen, L., Noponen, T., Meriläinen, P., et al. (2010). Hemodynamic responses to speech and music in newborn infants. *Human Brain Mapping*, *31*(4), 595–603.

Kudo, N., Nonaka, Y., Mizuno, N., Mizuno, K., & Okanoya, K. (2011). On-line statistical segmentation of a non-speech auditory stream in neonates as demonstrated by event-related brain potentials. *Developmental Science*, *14*(5), 1100–1106.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*, 684–686.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceeding of the National Academy of Sciences of the USA*, *100*(15), 9096–9101.

Ladefoged, P. (1975). *A course in Phonetics*. New York: Harcourt Brace Jovanovich.

Lennertz, T. J. (2010). *People's knowledge of phonological universals: Evidence from fricatives and stops*. Unpublished doctoral dissertation, Northeastern University.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329.

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*(2), 241–246.

Lin, C.-T., Huang, K.-C., Chao, C.-F., Chen, J.-A., Chiu, T.-W., Ko, L.-W., et al. (2010). Tonic and phasic eeg and behavioral changes induced by arousing feedback. *NeuroImage*, *52*(2), 633–642.

Lindell, A. K. (2006). In your right mind: Right hemisphere contributions to language processing and production. *Neuropsychology Review*, *16*(3), 131–148.

Lisman, J. (2010). Working memory: The importance of theta and gamma oscillations. *Current Biology*, *20*(11), R490–R492.

Lloyd-Fox, S., Blasi, A., & Elwell, C. E. (2010). Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy. *Neuroscience & Biobehavioral Reviews*, *34*(3), 269–284.

MacKenzie, H., Curtin, S., & Graham, S. A. (2012). 12-month-olds' phonotactic knowledge guides their word-object mappings. *Child Development*, *83*(4), 1129–1136.

Maguire, M. J., Brier, M. R., & Ferree, T. C. (2010). Eeg theta and alpha responses reveal qualitative differences in processing taxonomic versus thematic semantic relationships. *Brain & Language*, *114*(1), 16–25.

Makeig, S., & Jung, T.-P. (1996). Tonic, phasic, and transient eeg correlates of auditory awareness in drowsiness. *Cognitive Brain Research*, *4*(1), 15–25.

Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology*, *19*(23), 1994–1997.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, *6*(5), 314–317.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190.

May, L., Byers-Heinlein, K., Gervain, J., & Werker, J. F. (2011). Language and the newborn brain: Does prenatal language experience shape the neonate neural response to speech? *Frontiers in Psychology (Language Sciences)*, *2*, 222. (doi: 10.3389/fpsyg.2011.00222)

Mayberry, R. I., & Eichen, E. B. (1991). The long-lasting advantage of learning sign language in childhood: Another look at the critical period for language acquisition. *Journal of Memory and Language*, *30*(4), 486–512.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

McNealy, K., Mazziotta, J. C., & Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *The Journal of Neuroscience*, *26*(29), 7629–7639.

McNealy, K., Mazziotta, J. C., & Dapretto, M. (2009). The neural basis of speech parsing in children and adults. *Developmental Science*, *13*(2), 385–406.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*, 298–305.

Mikutta, C., Altorfer, A., Strik, W., & Koenig, T. (2012). Emotions, arousal, and frontal alpha rhythm asymmetry during Beethoven's 5th Symphony. *Brain Topography*, *25*(4), 423–430.

Milner, B., Branch, C., & Rasmussen, T. (1964). Observations on cerebral dominance. In A. V. S. de Reuck & M. O'Connor (Eds.), *Ciba foundation symposium on disorders of language* (pp. 200–214). London: Churchill.

Minati, L., Kress, I. U., Visani, E., Medford, N., & Critchley, H. D. (2011). Intra- and extra-cranial effects of transient blood pressure changes on brain near-infrared spectroscopy (NIRS) measurements. *Journal of Neuroscience Methods*, *197*(2), 283–288.

Molfese, D. L., & Molfese, V. J. (1979). Hemisphere and stimulus differences as reflected in the cortical responses of newborn infants to speech stimuli. *Developmental Psychology*, *15*(5), 505–511.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*(4), 495–500.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61–64.

Morgan, J. L. (1994). Converging measures of speech segmentation in preverbal infants. *Infant Behavior and Development*, *17*, 389–403.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66*(4), 911–936.

Mölle, M., Marshall, L., Fehm, H. L., & Born, J. (2002). Eeg theta synchronization conjoined with alpha desynchronization indicate intentional encoding. *European Journal of Neuroscience*, *15*(5), 923–928.

Nakano, T., Homae, F., Watanabe, H., & Taga, G. (2008). Anticipatory cortical activation precedes auditory events in sleeping infants. *PLoS ONE*, *3*(12), e3912. (doi: 10.1371/journal.pone.0003912)

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*(1), 1–19.

Nespor, M. (1993). *Fonologia*. Bologna: Il Mulino.

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-timed vs. syllable-timed languages. In M. Van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (Vol. 2, pp. 1147–1159). United Kingdom: Wiley-Blackwell.

Newport, E. L. (2002). Critical periods in language development. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (pp. 737–740). London: Macmillan Publishers Ltd./Nature Publishing Group.

Nyhus, E., & Curran, T. (2010). Functional role of gamma and theta oscillations in episodic memory. *Neuroscience & Biobehavioral Reviews*, *34*(7), 1023–1035.

Ockleford, E. M., Vince, M. A., Layton, C., & Reader, M. R. (1988). Responses of neonates to parents' and others' voices. *Early Human Development*, *18*(1), 27–36.

Ohala, J. J., & Kawasaki-Fukumori, H. (1997). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In S. Eliasson & E. H. Jahr (Eds.), *Language and its ecology: Essays in memory of Einar Haugen. Trends in Linguistics. Studies and Monographs* (Vol. 100, pp. 343–365). Berlin: Mouton de Gruyter.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Ordin, M., & Nespor, M. (under review). Lexical stress and phrasal prosody in segmentation. *Journal of Phonetics*. (initial submission in July 2012)

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*(2), 258–278.

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 80–96.

Papachristos, E. B., & Gallistel, C. R. (2006). Autoshaped head poking in the mouse: A quantitative analysis of the learning curve. *Journal of the Experimental Analysis of Behavior*, *85*(3), 293–308.

Parker, S. G. (2002). *Quantifying the sonority hierarchy*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

Parker, S. G. (2008). Sound level protrusions as physical correlates of sonority. *Journal of Phonetics*, *36*, 55–90.

Patel, A. D. (2007). *Music, language, and the brain*. New York: Oxford University Press.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674–685.

Peperkamp, S. (2007). Do we have innate knowledge about phonological markedness? Comments on Berent, Steriade, Lennertz, and Vaknin. *Cognition*, *104*(3), 631–637.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*, 604–607.

Peña, M., Maki, A., Kovačić, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., et al. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceeding of the National Academy of Sciences of the USA*, *100*(20), 11702–11705.

Peña, M., Pittaluga, E., & Mehler, J. (2010). Language acquisition in premature and full-term infants. *Proceeding of the National Academy of Sciences of the USA*, *107*(8), 3823–3828.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it?

Cognition, *95*(2), 201–236.

Pope, K. J., Fitzgibbon, S. P., Lewis, T. W., Whitham, E. M., & Willoughby, J. O. (2009). Relation of gamma oscillations in scalp recordings to muscular activity. *Brain Topography*, *22*(1), 13–17.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceeding of the National Academy of Sciences of the USA*, *103*(20), 7865–7870.

Querleu, D., Renard, X., Versyp, F., Paris-Delrue, L., & Crèpin, G. (1988). Fetal hearing. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, *28*(3), 191–212.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, *2*(1), 85–115.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292.

Reali, F., & Christiansen, M. H. (2009). On the necessity of an interdisciplinary approach to Language Universals. In M. H. Christiansen, C. Collins, & S. Edelman (Eds.), *Language Universals* (pp. 266–277). New York: Oxford University Press.

Rodríguez-Fornells, A., Cunillera, T., Mestres-Missé, A., & de Diego Balaguer, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society B (Biological Sciences)*, *364*(1536), 3711–3735.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105.

Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience*, *5*(7), 700–703.

Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, *9*, 295–302.

Sederberg, P. B., Kahana, M. J., Howard, M. W., Donner, E. J., & Madsen, J. R. (2003). Theta and gamma oscillations during encoding predict subsequent recall. *The Journal of Neuroscience*, *23*(34), 10809–10814.

Sederberg, P. B., Schulze-Bonhage, A., Madsen, J. R., Bromfield, E. B., Litt, B., Brandt, A., et al. (2007). Gamma oscillations distinguish true from false memories. *Psychological Science*, *18*(11), 927–932.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1–32.

Sievers, E. (1876/1893). *Grundzüge der phonetik zur einführung in das studium der lautlehre der indogermanischen sprachen*. Leipzig: Breitkopf & Härtel.

Singh, L., Reznick, J. S., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: A longitudinal analysis. *Developmental Science*, *15*(4), 482–495.

Singh, L., White, K. S., & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, *4*(2), 157–178.

Sirotin, Y. B., & Das, A. (2009). Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. *Nature*, *457*, 475–479.

Swingley, D. (2005a). 11-month-olds' knowledge of how familiar words sound. *Developmental Science*, *8*(5), 432–443.

Swingley, D. (2005b). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132.

Taga, G., & Asakawa, K. (2007). Selectivity and localization of cortical response to auditory and visual stimulation in awake infants aged 2 to 4 months. *NeuroImage*, *36*(4), 1246–1252.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*, 21.

Teinonen, T., & Huotilainen, M. (2012). Implicit segmentation of a stream of syllables based on transitional probabilities: An MEG study. *Journal of Psycholinguistics Research*, *41*(1), 71–82.

Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., et al. (2009). Sensitivity of newborn auditory cortex to the temporal structure of sounds. *The Journal of Neuroscience*, *29*(47), 14726–14733.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175.

Toro, J. M., Sebastián-Gallés, N., & Mattys, S. L. (2009). The role of perceptual salience during the segmentation of connected speech. *European Journal of Cognitive Psychology*, *21*(5), 786–800.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25–B34.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945.

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, *126*(1), 367–376.

Vallesi, A. (2012). Organisation of executive functions: Hemispheric aymmetries. *Journal of Cognitive Psychology*, *24*(4), 367–386.

Vallesi, A., & Shallice, T. (2007). Developmental dissociations of preparation over time: Deconstructing the variable foreperiod phenomena. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1377–1388.

Vouloumanos, A., & Werker, J. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159–171.

Watanabe, H., Homae, F., & Taga, G. (2011). Activation and deactivation in response to visual stimulation in the occipital cortex of 6-month-old human infants. *Developmental Psychobiology*, *54*(1), 1–15.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, *66*(4), 665–679.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, *7*, 30. (doi: 10.1186/1744-9081-7-30)

Witelson, S. F., & Pallie, W. (1973). Left hemisphere specialization for language in the newborn. Neuroanatomical evidence of asymmetry. *Brain*, *96*(3), 641–646.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46.