Multivariate pattern analyses of electrophysiological recordings in infants reveal the initial codes used by the human brain to represent speech, tones and number

Giulia Gennari

Université Pierre et Marie Curie

École Doctorale Cerveau Cognition et Comportment

Inserm-CEA Unicog, NeuroSpin

MULTIVARIATE PATTERN ANALYSES OF ELECTROPHYSIOLOGICAL RECORDINGS IN INFANTS REVEAL THE INITIAL CODES USED BY THE HUMAN BRAIN TO REPRESENT SPEECH, TONES AND NUMBER

Giulia GENNARI

Dissertation submitted for the degree of

Doctor of Philosophy in Cognitive Neuroscience

Supervised by Ghislaine DEHAENE-LAMBERTZ

Jury:

Richard ASLIN	Yale University	Reviewer
Teodora GLIGA	University of East Anglia	Reviewer
Christian BÉNAR	Aix-Marseille University	Examiner
Claire SERGENT	INCC- Université de Paris	Examiner
Ghislaine DEHAENE-LAMBERTZ	INSERM-CEA Unicog	Supervisor

[intentionally blank]

Acknowledgments

I had the luxury to be supervised by Ghislaine Dehaene-Lambertz. She welcomed me in her lab when I could not stick two lines of code together and knew barely nothing on infants and EEG. She is a deeply curious and truly insightful scientist, to an extent that I could have not imagined. Not only, she is the kind of supervisor that does not care for you just as a student, she regards you as a daughter. Thank you for being so incredibly inspirational and for believing in me way beyond what I could myself. Thank you for being there in the most difficult times.

The second great blessing I was given was to meet Sebastien Marti. He was the nicest mentor I could ever whished for. His advices and optimism kept being with me and helping me even after his disappearance. He is my role model. I miss him to this day, every day.

A huge *thank you* goes to some members of the developmental neuroimaging team. Merci to François Leroy, who was always there to provide any kind of help he could. Merci to Marie Palu, the energetic team manager, and Chanel Valera, the sweetest research assistant, because without them I would have simply had no data to analyze. Thank you to Ana Fló, for being such a kind colleague and for triggering my admiration. Thank you Milad Ekramnia for your cheerfulness.

I absolutely need to acknowledge two principal investigators at Unicog: Stanislas Deheane and Christophe Pallier. Thank you for making time in your crazy-busy schedule to listen and provide precious feedback.
I am also thankful to other members of the lab: Antonio Moreno and Isabelle Denghien for their IT assistance, Yair Lakretz and Evelyn Eger for important tips, and Vanna Santoro for being an awesome lab manager.
Grazie to Bianca Trovò and Fosca Al Roumi, who shared with me way more than office 4021.

I am grateful to the communities behind *stackoverflow.com*, *MNE Python* and *scikit-learn.org*, because, in full honesty, they were a major daily source of clarifications and resources.

Grazie to my parents for their invaluable support, especially throughout the last two years. Grazie to my grandma for her prayers and to my sister for her acceptance.

Last but clearly not least, the deepest *Merci* goes to Romain Ligneul; Dr. Didier Bouscary, Dr. Patricia Franchi and the staff at hôpital Cohin; Dr. Youlia Kirova and her team at Institut Curie. Without them I would have not had the opportunity to *live* this doctoral journey.

[intentionally blank]

ABSTRACT

Characterizing the origins of human cognition is among the most critical quests that scholars have undertaken since centuries. The work presented in this dissertation stems from the intuition that delineating the *initial set of representational primitives* engaged by our brain at the beginning of life can provide relevant pieces of understanding to this puzzle. Previous investigations of infant perception have relied on habituation and oddball paradigms, thereby tackling *discriminative skills* rather than encoding mechanisms in their essence. Here, we combined high-density electroencephalography and multivariate pattern-analysis techniques to characterize how 3-month-old infants encode speech, numerical quantity and musical pitch. The capability of multivariate-pattern analysis to grasp macropopulation codes enabled us to gain direct access into the informational content represented by infant brains within ecological settings. In each study, we provide novel insights to solve long-lasting debates. To start with, whereas many researchers have rejected the possibility of authentic phonetic processing during the first semester, we document how the infant brain decomposes the speech input along orthogonal minimal dimensions, corresponding indeed to the phonetic features described by linguists. Second, during the last couple of decades, there has been a heated debate regarding the roots of numerical competence: classical theories have been challenged by the idea that a generalized magnitude system, conveying all sensory streams into holistic estimates, might correspond to the only type of quantification available in infancy. Oppositely, we show that 3-month-old brains extrapolate approximate numbers separately from non-numerical correlated variables and in a completely automatic manner. Strikingly, we could demonstrate the existence of an extraction mechanism transcends sensory modality, presentation format and wakefulness state, revealing a genuinely abstract numerical code. Lastly, we describe how young infants process musical pitch along its two psychological dimensions: height and chroma. Whereas the latter parameter is often considered as a higher-order product of Western culture, we demonstrate how, instead, pitch chroma corresponds to a basic organizing principle of neural responsivity that is observable very early in development. Overall, we have disclosed primitive encoding systems that are particularly advantageous for at least three reasons: they alleviate the burden of input dimensionality by compressing it; they overcome sensory variability by capturing invariance; they depict aspects of the world that are highly relevant, and thus adaptive, for the human being. Given these strategic characteristics, their early onset and the representational flexibility they afford, we believe that neural codes for phonetic features, approximate number and pitch quality provide an ideal ground for knowledge acquisition and can thus serve as catalyzers for the development of human cognition.

Table of Contents

List of Figures

Chapter 1. GENERAL INTRODUCTION

Starting from the first second of life, our brain receives a flow of sensory inputs that is staggering in its complexity. With 30,000 auditory fibers and 106 optic fibers (Tenenbaum et al., 2000), our nervous system transmits inputs that are not just impressively high-dimensional but also ever-changing. As a matter of fact, although we typically see an object and hear a word many times, we effectively never encounter the same image or the same sound twice. Nevertheless, adult human beings come to possess a rich understanding of the world, an essential prerequisite for the functioning and well-being of both individuals and communities. Such a refined knowledge is obtained, flexibly exploited and constantly updated by means of a heterogeneous and well-organized system of mental abilities – our cognitive functions. How does the infant brain deal with the chaotic, overly complex sensory input? How can humans achieve, in just a few years, the optimal level of functionality that characterizes adult cognition?

Comprehending the genesis of our mind is among the most critical quests philosophers and scientists have undertaken since centuries. Where shall we start in this strive? An idea common to modern psychological theories is that cognition depends on specialized structures that take the form of *representational* items, corresponding to internally-processed bodies of information somewhat akin to theories (Gopnik & Meltzoff, 1998; Carey & Spelke, 1994; Fodor, 1983). In philosophy of mind, classical computational theories conceive cognition as a system of algorithmically specifiable operations defined over structured mental *representations* (Horst, 2003). Thus, irrespective of the epistemological approach embraced, it would seem that any theory of cognitive development must begin with delineating the **initial stock of representational primitives**[1] available to humans at the beginning of their life.

Our experiential world can be characterized at multiple representational levels. For instance, we are able to perceive a basket of peaches placed right in front of us because a precise pattern of light, at a given moment, hits our retina. According to a first level of description our basket of peaches is nothing else than a series of spatially distributed luminance values. In such a pattern, oriented contrasts define edges and curvatures, separating the fruit from the background. Next, the peaches can be said to occupy a certain area and have a certain weight. Now, we may change our own position or that of the basket, we may close the blinds and switch on a lamp; the descriptors referring to the volume and the weight of the peaches will remain valid despite prominent changes in brightness and contrasts. There are more-or-less a dozen of peaches in the basket. When next week these twelve

[1] Although slightly out of the original context, this expression is borrowed from the book *The Origins of Concepts* by Susan Carey (2009).

pieces will be replaced by a fresh dozen, the basket will contain peaches occupying a different volume and weighting a different amount of grams, creating a different pattern of edges and curvatures and, still, there will be twelve peaches in front of us. Even when their number changes, as we would really love a fruit salad right now, we will remain able to recognize those in front of us as members of the category "peach", irrespective of changes in brightness and contrast, in volume, in weight and in numerosity. This naïve, perhaps fuzzy, example demonstrates how, as adults, we can encode the external world by means of various descriptors. Crucially, each descriptor can be generalized (only) to a particular extent, thereby affording distinct levels of representational flexibility.

The experimental work presented in this thesis stems from the intuition that delineating the initial descriptors used spontaneously and pre-attentively by the infant brain to encode the external world might provide meaningful insights upon the origins of our cognition.

1.1. Theoretical and terminological premises

A basic idea permeating the entire dissertation is the conception of the brain as an adaptive information-processing device: energy patterns coming from the environment and captured by our senses are processed precisely within the scope of extracting information that is useful for the organism. Within this perspective, understanding the preliminaries of human representations means identifying *what* information is processed early in life and, crucially, *in what format*.

With the term *representation* we assume a relatively lightweight notion that is shared by most neuroscientific works. Specifically, a representation is any internal state of a complex system that serves as a vehicle for **informational content** and plays a functional role within the system based on the information that is carried (Bechtel, 1998). When mentioning *neural codes* we refer precisely to this description: a pattern of cortical activity, i.e. an internal state of the brain, that *encodes* information, i.e. represents informational content, for later usage (deCharms & Zador, 2000). Whereas the concept of representation entails both a content and its function, the experimental work presented in this dissertation is limited in this respect (Ritchie et al., 2019), as we probed the content and properties (e.g. temporal dynamics) of neural representations directly but allude to function inferentially.

Note that we could bypass representational interpretations altogether and approach the brain as a dynamical system (Bechtel, 1998). Such a perspective focuses on physical mechanisms and, ultimately, would account for all aspects of brain activity. Nevertheless, the notions of information/representation can help us understand neuronal dynamics at a broader level[2].

[2] To better express how that is the case we could make an analogy with computers: they "can be understood as dynamical systems. However, interpreting the patterns of charges and currents as representations of data and instructions enables us to capture a computer's behavior more concisely in a high-level algorithmic description that reveals the dynamics in terms of the implemented functions" (Kriegeskorte & Diedrichsen, 2019).

With the adjectives *infant* and *initial* we allude to the time period prior the 6th postnatal month; this choice is motivated by the fact that, as it will be explained in more detail throughout the dissertation, the second semester has been proposed to coincide with a turning point in human brain development, both structurally and functionally.

1.2. Methods to investigate infant representational skills: the state of the art

The assessment of mental processes during the very first moths of life is challenged by extreme limitations such as the impossibility to deliver verbal instructions or collect verbal reports, a very narrow motor repertoire, low cooperation and tolerance during experimental sessions. Nonetheless, developmental researchers have been able to provide substantial contributions to shed light into the origins of adult representations by studying this population.

1.2.1. Behavioral measures and the habituation paradigm

Major advances were first made possible by capitalizing on two simple behaviors, visual fixation and sucking, and the spontaneous tendency of all animals to decrease reactivity to repeated stimuli but re-engage their interest when the stimulus changes (Schöner & Thelen, 2006). The exploitation of these elements led to the emergence of the habituation/familiarization paradigms, which are currently the most common expedient used to study infant cognition. Within a classical habituation study, subjects are repetitively exposed to the same items (or class of items) until their looking time or sucking rate declines. Subsequently, they are tested with new exemplars where the characteristic under study is varied. A recovery of fixation or sucking rate is interpreted as evidence that the infant detected the difference between the habituating displays and the test stimuli, thereby proving to be capable of representing a given characteristic. This kind of paradigms has led to tremendous advancements in our knowledge of early representations. For instance, they have revealed that since their first days of life infants can discriminate speech tokens that differ along subtle, linguistic-relevant dimensions such as voice onset time, manner and place of articulation. Just like adults, young infants seem to perceive syllables categorically along these dimensions and to normalize across the acoustic variations that derive from speaking rate and voice peculiarities (for an extensive review, see: Jusczyk, 2000). Whereas in the early 60s mainstream ophthalmology textbooks claimed that newborns were blind, habituation studies have shown that they already recognize an object based on its shape or its size across retinal changes (Slater & Morison, 1985; Slater et al., 1990) and they can distinguish visual arrays containing two versus three dots (Antell & Keating, 1983).

Although being able to provide crucial insights as those just listed, the employment of behavioral dependent measures to infer hidden representational states poses important interpretational dilemmas. The latter arise from the fact that these studies entail a many-to-one mapping problem: numerous (alternative or complementary) factors contribute to a single dependent variable (Aslin,

2007). To start with, the possibility for the infant to dishabituate requires the subject to *remember* the characteristics of the repeating stimulus and eventually *compare* stored items with the new input. This means that discrimination and memory are always confounded within the results. Another interpretational limitation concerns the frequent occurrence of familiarity effects: within the test phase, young subjects might happen to show increased looking time or sucking rate for repeated instances instead of novel stimuli. Such a phenomenon is due to the fact that infant preference is naturally driven by both novelty *and* familiarity (Oakes, 2010). Given the prominent difficulties inherent to developmental studies, any significant effect is normally interpreted as evidence for discrimination (e.g. Cantrell & Smith, 2013). Specifically, dishabituation to novel conditions can be considered an attentional response (e.g. re-orienting; Sokolov, 1963), while familiarity effects are assumed to reflect recognition (Aslin, 2007). Beyond its ambiguity per se, such a practice has the serious downside of complicating the integration of results coming from different investigations. Crucially, bidirectional effects might cancel each other out (i.e. the two opposite preferences might counterbalance each other) such that null results can indicate a true impossibility to discriminate as well as a false negative. Memory-related factors and familiarity are perhaps the major, but certainly not the only sources of confounds. Just to mention a few others (that are not avoidable by means of a careful experimental design and rigorous data collection), arousal state, attentional disengagement, blank stares or cross-modal competition have all been reported to influence dishabituation performance (Colombo & Mitchell, 2009; Aslin, 2007).

1.2.2. Neural measures and the oddball paradigm

The advent of functional neuroimaging enabled to overcome some of these interpretational dilemmas, as it provided the tremendous practical advantage of not requiring an overt response. Moreover, it made possible to investigate early representations from a completely novel perspective: that of functional neural architectures. Among the techniques currently available, electroencephalography (EEG) is the best suited for developmental research[3]: it is the least invasive; it requires a simple recording system; and is less negatively impacted by movements. Most commonly, research on infant cognition focuses on the event-related-potentials (ERPs), consisting of changes in voltage that are time-locked to the experimental manipulation (e.g. stimulus-evoked). Scalp ERPs derive from the activation of groups of cerebral neurons[4] that are both synchronized and spatially aligned such that to generate electrical fields recordable at the surface of the head. Intriguingly, whereas behavioral measures reflect the final product of many intermixed mental

[3] Other promising techniques for functional brain imaging in infants include Magnetoencephalography (MEG; e.g. Kujala et al., 2004) and functional Near-Infrared Spectroscopy (fNIRs; e.g. Lloyd-Fox et al., 2010). Yet, to date, technical progresses are still needed in order to optimize their suitability for this delicate field.
[4] more precisely: the postsynaptic depolarization of cell dendrites (Luck, 2014)

operations (which, as we have seen, are very hard to disentangle), EEG offers the opportunity to observe brain processes at play in a time-resolved fashion.

To target early representations, the vast majority of neuroscientific studies conducted so far has employed oddball paradigms (Näätänen et al., 1978). Within the latter, a repetitive stimulus (called standard) is occasionally exchanged with another item differing in a feature of interest or violating an expected pattern. By comparing the brain waves elicited by the repeated stimuli to those triggered by deviants it is possible to isolate mismatch responses (MMRs), which are regarded as reflecting the detection of a change. Whereas the logic behind behavioral habituation and oddball paradigms is essentially the same, mismatch responses can be elicited irrespective of where infants focus their attention and even when they are asleep (Cheour et al., 2000), a phenomenon that makes this method incredibly versatile. Charmingly, this methodology revealed that early representational skills are supported by an (already) intricate functional organization within the infant brain composed of specialized modules and parallel pathways. For instance, when two different dimensions are contrasted within the experimental paradigm, such as the talker's voice and the linguistic value of spoken syllables, or the identity[5] and the number of objects in visual displays, the mismatch responses of 3-month-olds are characterized by similar latencies but different topographies (Bristow et al., 2008; Izard et al., 2008). In this context, distinct topographical distributions of neural activity evidence the engagement of different sources, indicating that the dimensions under investigation are processed, in parallel, by separate neural networks (Dehaene-Lambertz & Spelke, 2015; Dehaene-Lambertz & Gliga, 2004).

Although powerful, the use of MMRs to investigate neural representations does not come without important drawbacks. First, these responses require the presence of short-term memory traces for the repetitive aspects of the stimuli, together with some form of (even if extremely rudimentary) comparison mechanism (Winkler, 2007). Thus, just as behavioral measures, this methodology fails at targeting pure encoding mechanisms. Considering that the capacity of short-term memory is extremely limited early in development (Ross-sheehy et al., 2003), this might be a crucial limitation: some aspects of the incoming stimuli might be processed and yet not storable.
The second important limitation pertains to the fact that cross-individual variations in the amplitude and latencies of the ERPs are broad, considerably larger than what observed for adults (e.g. Cheour, Alho, et al., 1998; Naik et al., 2021). Such variability arises, in first place, from heterochronocity: although the sequence of developmental changes in the shape of the mismatch response may be fixed (to a certain extent), the exact rate of change varies from subject to subject (Courchesne, 1990). Another crucial source of variation are logistical constraints: has the reader ever tried to place an EEG cap exactly in the same position over the scalp of ~50 babies who are laughing, crying, about to cry, turning their head all over the place or trying to chew the sensors in the meanwhile? It is

[5] defined by a combination of shape and color

practically impossible for one electrode to cover the same exact scalp position in all, or even a third, of the participants[6]. Inter-subject variability is worrisome because data analysis is performed **across participants**: when MMRs from different subjects are pooled together in order to test for significant effects, the peculiarities of each individual response are likely to cause signal loss, at the detriment of methodological sensitivity. Other than reducing our chance of uncovering brain function, inter-subject variability prevents us to detect dysfunction: since normative values cannot be reliably estimated, MMRs are ill-suited when it comes to discern whether some representational skills are missing or aberrant (Picton & Taylor, 2007). Although not directly related to the scope of the present thesis, this impossibility is crucial, as the purpose of neuroscience research is not only to understand brain function but also to diagnose, prevent and cure its potential malfunctioning.

1.3. Infant representational primitives are yet to be defined

Despite suboptimal sensitivity, behavioral and neuroimaging studies have demonstrated the existence of quite impressive **discriminative** skills within the first semester of life. Yet, the initial units used by the human brain to encode the external world remain to be discovered: what information do infants use in order to discriminate among sounds or images?

Several researchers and theoreticians propose major differences between early and mature representational units. Starting from the linguistic domain, many would argue that the ability of young infants to differentiate between spoken syllables relies on a refined acoustical analysis of speech rather than adult-like phonetic units (Kuhl, 2004; Vilain et al., 2019). Switching to mathematics, the possibility to discern between numerical arrays displayed by humans since their first hours of life might be based on the encoding of alternative magnitudes that inevitably correlate with number (Mix et al., 2002). Some have proposed quantitative processing to be "one-bit" during the first months (Walsh, 2003) such that babies might be able to differentiate areas, lengths, numbers or durations based on a unique and generalized code for "size" or "more/less" (Leibovich et al., 2017; Hamamouche & Cordes, 2019). Even in the domain of pitch perception, prior to the 4-5th month of life, electrophysiological observations seem to suggest that infants might distinguish musical tones by relying on plain, isolated frequency components rather than adult-like integrated percepts (He & Trainor, 2009).

Permeating these proposals is the idea that initial encoding units may be somewhat more rudimentary relative to those employed by mature brains: depending on the particular domain at hand, it is often assumed or implied that early representational units are more closely related to the physical input, holistic rather than fine-grained, or fragmentary rather than integrated. Conceptions of this kind are reasonable and plausible, not only in light of inexperience but also on the basis of

[6] This difficulty has not been encountered solely by the student who writes, as more formally documented by Kabdebon et al., (2014).

anatomo-structural considerations. Taking the auditory system as a reference, the maturation of primary auditory cortex and acoustic radiations is far from being complete at birth; conversely, it extends until the third year (Yakovlev & Lecours, 1967). Whereas in the adult brain 98% of the sensory input is transmitted through thalamocortical pathways that synapse in cortical layer IV, axonal maturation is restricted to cortical layer I during the first ~20 postnatal weeks (Eggermont & Moore, 2012). Further, the myelination of temporal regions starts only after the fifth month (Pujol et al., 2006). Since thalamocortical connections remain severely immature during the first semester, behavioral and evoked responses to sounds are likely to be mediated by quite peculiar sensory pathways early in life (Werner et al., 2012). For instance, they may comprise a role for the subplate, a transient structure that hosts thalamic afferents prior maturation of the cortical plate (Luhmann et al., 2018; Molnár et al., 2020; Wess et al., 2017). It is fascinating in this regard that, despite structural immaturity, the gross pattern of cortical regional activation triggered by music in neonates (Perani et al., 2010) and by spoken sentences in 3-mo-olds (Dehaene-Lambertz, 2002) is remarkably similar to that seen in older children and adults.

Although being able to bring invaluable insights upon early discriminative abilities, neural architecture and functional specialization, current methods are poorly suited to delineate representational primitives in that they leave primary encoding processes underspecified.
To express how that is the case, let us take as an example phonetic discrimination. Cheour and colleagues (1998) have investigated the mismatch responses elicited by the vowels /õ/ and /ö/ in 6-mo and 1-y old infants who were exposed to many repetitions of the vowel /e/. Crucial for their paradigm the fact that /õ/ is acoustically more dissimilar form the standard /e/ relatively to /ö/ but, unlike the latter, does not have any phonemic value in Finnish (while it does within other languages, e.g. Estonian). The MMR observed in 6-mo-old Finnish infants was comparable across the two deviant conditions, whereas in 1-year-olds the vowel belonging to the Finnish repertoire elicited a greater mismatch (despite its lower acoustic deviancy). With this set of results, the authors have demonstrated the development of "language-specific memory traces" within the second semester of life (Cheour, Ceponiene, et al., 1998). In what format did the younger participants encode the experimental stimuli? A spectral code (based on e.g. formant frequencies) and a phonetic code (based on e.g. tongue height and/or backness[7]) can explain the observation of a mismatch response equally well. As a matter of fact, whereas some researchers have considered evidence of this sort as indicating domain-general spectrotemporal processing of speech, these results are equally consistent with the idea that the brain learns, via exposure, to bypass or down-weight irrelevant phonetic distinctions (i.e. those not belonging to the repertoire of the mother tongue) while keeping track of a (linguistic) repetition.

[7] This terminology is borrowed from the classical nomenclature used in linguistics, without alluding to any particular theory.

More broadly speaking, by manipulating what characteristic of the stimulus elicits a mismatch response it is possible to infer whether a neural network detects a regularity in the dimension under study. As noted above, to achieve this task, the system *needs* to retain a certain feature of the stimulus. Although undoubtedly meaningful, the observations provided by this method (as well as by any experimental paradigm based on change-detection) do not speak to the representational units used by the system to encode the stimuli. For instance, in a real word scenario[8], the network might compute a certain code in preparation of an immediately subsequent processing stage without keeping a record of it. Moreover, even when the nature of the experimental design is taken scrupulously into account, interpretational ambiguities remain often inevitable. For instance, as explained more extensively in Chapter 3, when manipulating what counts or not as a repetition, it is physically impossible to separate numerical information from correlated non-numerical features, preventing to discern what kind of quantitative dimension the system discriminates.

1.4. Approaches to data analysis: univariate versus multivariate

Overall, the developmental neuroimaging studies conducted so far to investigate early representational skills have relied on the so-called univariate analysis approach. Within the latter, measurements derived from each recording channel are treated as independent pieces of data, typically using statistical tests to determine whether experimental conditions triggered different deflections in the neural signal. The term *univariate* refers precisely to the fact that within this approach the analysis of one channel has no impact on that of any other. Often, neural recordings have been collected from just a handful of scalp sites; alternatively, neural signal has been pooled across neighboring locations in order to improve the signal-to-noise ratio and mitigate the problem of multiple comparisons.

Yet, it is nowadays well established that individual pieces of information are carried not by single cells but rather by neuronal ensembles, a strategy known as population coding (Pouget et al., 2000). A population code is a complex set of activity patterns that cannot be satisfactorily grasped by a simple average or summation of the signal coming from neurons within a circumscribed ensemble (deCharms & Zador, 2000; Jacobs et al., 2009). In fact, a population code arises from the combination of diverse, complementary and synchronous factors such as the interplay between the coarse firing rate of localized groups of cells and the precisely-timed spike patterns of more distributed neurons that respond sparsely over time (Panzeri et al., 2015; Stanley, 2013). These observations suggest that, when the goal is to discover the information processed by the brain, considering each channel/voxel separately, as done in conventional analyses of developmental neuroimaging data, might be quite restrictive. That is to say, in addition to isolated activation (e.g. voltage) levels, analyzing the relative

[8] The reader shall be reminded of the fact that mismatch responses are not recorded as such from the scalp, they are an artificial construct of the investigator

differences in activity between channels could potentially provide a more comprehensive picture of brain function.

Crucially, seminal fMRI studies on adult visual perception have demonstrated that such a consideration holds even if the neuroimaging technique under employ allows assessing neural activity only at a macroscopic level. For instance, Haxby and colleagues (2001) have shown that despite overlapping activations within the ventral temporal lobe, the *patterns* of activity across this region are discriminative of object categories even when the most responsive voxels are ruled out from the analysis. At the time of the study, this was a groundbreaking finding as it highlighted that, other than modular organization and localized tuning, distributed combinatorial codes might play an important role in object recognition. Strikingly, a subsequent study on line orientation (Kamitani & Tong, 2005) showed that although selectivity for this parameter exists at a sub-voxel level (i.e. within narrow cortical columns at a scale of a few hundred micrometers) in early visual cortex, there are still small irregularities in the way the activity captured by each voxel reflects different orientations. When multiple voxels are jointly analyzed, discriminating perceived orientations becomes possible whereas such a task could be achieved only with the aid of invasive imaging techniques beforehand.

The approach adopted in the latter studies is referred to as *multivariate*. Generally speaking, multivariate analysis is a set of methodologies that take into account the relationship between multiple variables (instead of treating them as independent) in order to unravel patterns in the data. Within the neuroimaging field, these analyses are designed to test whether two or more experimental conditions can be distinguished on the basis of multiple measurements of neural activity, recorded during each experimental instance. When that is the case, it can be concluded that information pertinent to the manipulation of interest exists in the neuroimaging data (Mur et al., 2009). Whereas ignored by the univariate approach, multivariate analyses take into account the relative contribution of each data point to discriminability as well as their covariance. Such a strategy makes these methodologies a powerful tool, sensitive not only to regional-average activation differences but also to changes in granular patterns.

We propose that a switch from an univariate to a multivariate approach might be particularly fruitful within the quest of representational primitives for at least three reasons.

a. *Neural processes can be captured more exhaustively*

   Mainly, the promise of this approach relies in its neuroscientific rationale. As explained above, the brain operates by means of distributed, spatially extended macropopulation codes. In order to characterize these codes to the fullest the patterns of neural activity that are sparse and/or fine-grained need to be taken into account as they are themselves, among others, the holders of the message. Whereas univariate methods override these dimensions altogether, the multivariate approach is ideally suited for this task (Haynes & Rees, 2006).

b. The (highly dubious) assumption of shared topographies is no longer needed

Multivariate procedures are typically conducted **within subject**: to be pooled across participants is not the neural data itself but their projection onto psychological dimensions. On the technical side, such a strategy is clearly preferable and preferred in order to preserve the high-spatial-frequency patterns of response, which would be smoothed away otherwise. From an exquisitely neuroscientific perspective, such an expedient is appealing in that each brain is structurally and functionally unique and while we can assess or formulate hypotheses about which perceptual and cognitive dimensions are shared, the degree of inter-subject correspondency at the neural level remains unknown. Such a characteristic becomes even more advantageous in infant studies where idiosyncratic developmental trajectories translate to conspicuous inter-individual variability of structural and functional brain topographies.

c. Noise level is drastically reduced

By combining data from multiple sensors/voxels it is possible to cancel out the noise (Haufe et al., 2014), thereby maximizing the detectability of meaningful signal. Although purely technical, this feature is extremely appealing for the developmental neuroimaging field, where the signal-to-noise ratio in the data is usually very low. In particular, neural signals from young populations are more prominently contaminated by motion artifacts: infants tend to move often and abruptly and they cannot be provided with verbal instructions to refrain (Fujioka et al., 2011). In addition, physiological artifacts become more problematic, since the polygraphic measures monitored in adult studies, such as the electrocardiogram (ECG), the electromyogram (EMG) or electrooculogram (EOG), are too impractical to obtain. Concerning EEG, infant recordings are characterized by ample and irregular background activity that inevitably conceals the tiny evoked responses of interest (Bell & Wolfe, 2008).

1.5. Zoom on multivariate pattern classification

In the last decade or so, the use of machine learning algorithms to classify neural response patterns has become increasing popular in adult neuroimaging. This multivariate methodology is often referred to as "**decoding**", alluding to the fact that its goal is not to model information processing per se, but rather to reveal the content of the code.

A typical classification analysis begins with dividing the neural data into independent training and test sets. Machine learning algorithms consider each element in the dataset (e.g. each channel or voxel) as a separate dimension (or separate 'feature') in a high-dimensional space. When the recording system includes *N* voxels or channels, each stimulus presentation elicits a response vector that occupies a point in an *N*-dimensional neural activation space. Classifiers are fitted on the training set in order to find a decision boundary that can efficiently separate the response vectors associated with each experimental condition (Figure 1.1). The resulting model is evaluated by means of its

performance on the neural data that was left out from training: classifiers are used to "predict"[9] the experimental conditions characterizing the trials in the test set. Their performance in this task can be conceived as an estimate of the information about the experimental variables contained in the neural data. Specifically, if classifier performance is higher than that expected by chance it is assumed that the patterns of brain activation contain information that distinguishes between the experimental conditions. Note that relatively to univariate analyses, the use of an independent test set provides a nice statistical advantage: the assumptions of the model are implicitly tested when we assess its predictions (Kriegeskorte, 2011). If assumptions are violated performance will suffer, implying a minimal risk of false positives. Notably, out-of-samples estimates still require second-order statistical procedures, normally performed at the group level, in order to establish their reliability.
The reliability of the classification performance can depend on various factors such as the signal-to-noise ratio in the neural data and, mostly, on the number of trials available for the analysis, since only an adequate number of samples will allow the robust calculation of an optimal decision boundary. The common strategy used in adult studies to ensure reliable estimates is k-fold cross-validation: the neural data available is split in *k* groups[10] and the procedure of model training and out-of-sample prediction is repeated *k* times. At every run one fold is hold out from training such that within the entire loop each data sample is tested once and an overall measure of performance is obtained.

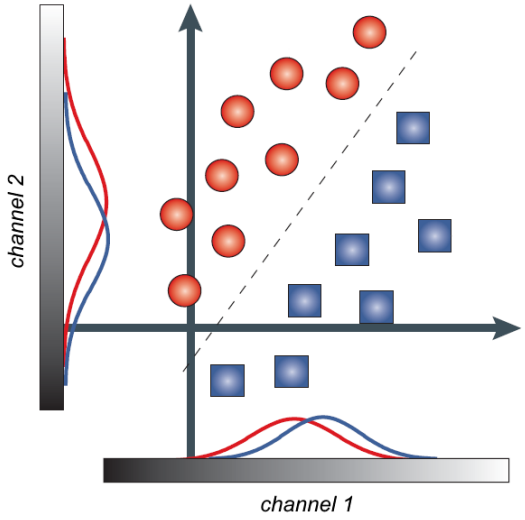

Figure 1.1 Linear decision boundary within a simplified neural activation space.
Red circles and blue squares illustrate the responses to two experimental conditions that cannot be separated from individual channels due to largely overlapping activity distributions. The responses are plotted in a two-dimensional space, corresponding to the activation captured by two channels at one time point. Crucially, by taking into account the combination of the responses from both channels it is possible to define a boundary (dashed line) that separates the two classes. Adapted from Haynes and Rees (2006).

[9] In neuroimaging "prediction" is used with a figurative connotation: this term denotes guessing the experimental condition (e.g. which stimulus was presented) from the neural activity pattern (Kriegeskorte & Bandettini, 2007). Of course, such an act of "prediction" (i.e. the analysis) occurs *after* the predicted events (i.e. the experimental session).
[10] Normally, *k* ranges from 3 to 10. Single trials are assigned to the groups on the basis of random partitioning or experimental (sub)conditions.

In order to achieve correct classifications, information must be present within the neural data in a format that the decoder can exploit. Most often, researches within the adult neuroimaging field have chosen to use **linear** models, requiring the distribution of patterns to be linearly separable to some extent. In linear classification, each feature of the high-dimensional space receives one weight parameter and the product between weight and response vector is used to assign class membership to the response patterns at test. Although this process can miss the presence of information encoded in a more complex manner, linear models are preferable for two reasons. First, they overfit less easily than nonlinear algorithms yielding more stable estimates. Most importantly, they facilitate interpretation as they capture only information that can be read out in a single biologically plausible step by a downstream brain region (Kriegeskorte, 2011). Conversely, since all linear classifiers fit a hyperplane to achieve class separation, the particular algorithm used (e.g. support vector machines, multiple regression, Fisher discriminant) has marginal relevance for the neuroscientific interpretation of the outcome.

1.5.1. And in babies, why not?

Multivariate pattern-analyses hold important promises for the investigation of human brain processing. Throughout the previous paragraphs, we have seen how such potentialities would be particularly precious to uncover the content of infant representations. Nevertheless, the employment of multivariate decoding within the developmental field has been precluded by prominent (practical) constraints. First and foremost, whereas multivariate decoding imperatively requires **large amount of data** samples in order to obtain reliable estimates, the experimental protocols used in infants have typically quite short data collection times due to reduced tolerance for testing. Another important difficulty concerns **intra-subject variability**: variations across single trials belonging to the same condition are consistently more prominent in infants relatively to adults (Coch & Gullick, 2012; Picton & Taylor, 2007). This might be partly due to frequent changes in arousal or attentional states, determining fluctuations in background activity. When that is the case, as outlined above, multivariate analysis might be optimally suited to detect and rule out this source of noise[11]. Definitely more problematic is the presence of fluctuations within the neural networks that are processing the stimuli, which creates inconsistencies across the actual ERPs (Thomas & Crow, 1994). Such a troublesome phenomenon might be caused by synaptic inefficiency (perhaps related to early overproduction) and axonal a-synchronicities, due to irregular myelination. These factors translate to inconsistent strength or timing of the neural firing patterns underlying the evoked responses. Overall, insufficient amount of data combined with such peculiar inter-trial variability is likely to prevent classifiers to achieve stable decision boundaries.

[11] Variability could also reflect a processing strategy of the brain (e.g. the event-related variability described by Naik et al., 2021). Interestingly, in this case, multivariate classifiers are designed to recognize its relevance in respect to the experimental variables and thus treat it as signal of interest.

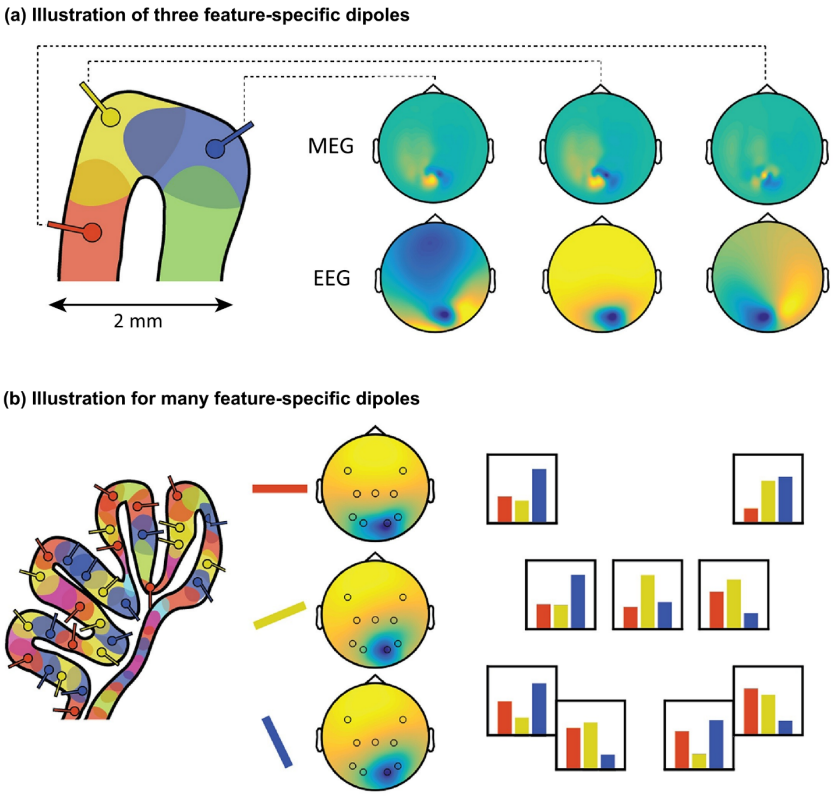

Figure 1.2 Variation in the angle of neighboring dipoles generates separable signals at the scalp surface.
(a) three dipoles approximately 2 mm apart but with very different angles result in easily distinguishable MEG (upper row) and EEG (lower row) topographies. (b) Such differences will tend to average out with increasing numbers of dipoles, resulting in very similar topographies. Crucially, multivariate pattern analysis can differentiate fine stimulus features by pooling the specific information contained in the subtle biases of each sensor. Adapted from (Stokes et al., 2015)

1.6. The spatial richness of non-invasive electrophysiological measures

Since multivariate methodologies capitalize on spatially-resolved details, the application of pattern-analysis was initially restricted to fMRI studies. This was justified by the classical dichotomy according to which non-invasive electrophysiological measures of brain activity (i.e. MEG and EEG) entail excellent temporal resolution at the expenses of spatial sharpness, whereas the contrary applies to hemodynamic measures. Such a conception stems from the fact that it is not possible to localize with certainty the anatomical sources of MEG/EEG signals. Yet, with an impressive set of experiments, Cichy and collaborators (2015) have shown that when the purpose is to track differential information rather than to localize activity differences, source ambiguity hardly matters. Namely, these authors demonstrated that MEG recordings support the decoding of fine visual elements processed at the level of cortical microstructures (Cichy et al., 2015) just as fMRI images (Kamitani & Tong, 2005), thereby revealing how the degree of spatial details embedded in MEG data had been largely underestimated.

Where does such spatial richness come from? Over the surface of the scalp, magnetoencephalography measures the magnetic field of dipoles generated by the electrical activity of spatially aligned cells (Luck, 2014). The distribution of these fields depends on the location of the dipoles and, critically, on their angle. Due to cortical surface irregularity, even dipoles from neighboring groups of cells will

have different angles, translating to separable field patterns (Figure 1.2). Provided an adequate number and disposition of sensors, the same mechanism holds true for the electric potentials captured by electroencephalography. It is intriguing in this context that, whereas adult EEG is affected by the smearing exerted by the scalp and the skull, these barriers present a markedly lower impedance in infants, resulting in an extremely rich spatial texture (Grieve et al., 2004; Odabaee et al., 2013). The opportunity to exploit such a peculiarity adds to all the advantages listed so far in delineating multivariate pattern analysis as an extremely promising approach for the study of early encoding primitives.

1.7. A framework for the present thesis: representational spaces (Haxby et al., 2014)

The goal of this thesis is to adapt some of the multivariate techniques previously employed on adult data in order to characterize the initial units used by the human brain to encode speech, numerical quantity and musical pitch. Starting from the current understating of both adult and infant perception, we created two main experimental paradigms where sets of relevant parameters were manipulated to create an informative **multidimensional space**. In the latter, each stimulus can be characterized by/conceived as a vector with different values along the distinct dimensions of interest.

Given the "imaging advantage" offered by anatomical immaturity and inspired by the application of multivariate analyses on ECoG data (Mesgarani et al., 2014), we strived at capturing fine-grained activity patterns as accurately as possible. In this attempt, we recorded evoked neural responses by means of a prototype super-high-density EEG system featuring clusters of electrodes distanced only 5mm one from the other (Figure 1.3).

In parallel with the characterization of the stimuli proposed above, also the neural activity elicited by our paradigms can be described as a multidimensional space composed of response vectors where each value is a measure of local activity (i.e. a voltage captured by a given EEG sensor at a given time). At their core, our investigations consisted in testing for relationships between the stimulus space, as pictured by the independent variables of interest, and the neural space, as captured by our EEG system.

Mainly, we employed pattern classification algorithms to test the presence of sectors in the neural multi-dimensional space in which all response vectors embed the same class of information. The possibility to reliably associate neural reposes to experimental conditions demonstrates the existence of a statistical dependency between the stimulus space and the neural space. Crucially, as outlined more in detail in the methodological sections of each chapter, such a demonstration was always insufficient to our goal. Striving to delineate the format of infant neural codes, we systematically resorted to cross-decoding: we trained algorithms on subparts of the stimulus space

and assessed their performance on alternative portions where a particular dimension was altered. To corroborate pattern-classifications we employed Representation Similarity Analysis (RSA), a multivariate technique that examines the structure within the neural space in terms of distances between response vectors. Specifically, this method tests whether the similarity between brain responses to different stimuli matches the similarity between the stimuli according to a specific model of representation. Since with RSA the variables manipulated within the experimental paradigm can be sampled more comprehensively, this approach entails the opportunity to forgo any predefined stimulus grouping (Kriegeskorte, 2011). Such a potentiality enabled us to probe the validity of the a-priori theoretical assumptions used to formulate our decoding problems.

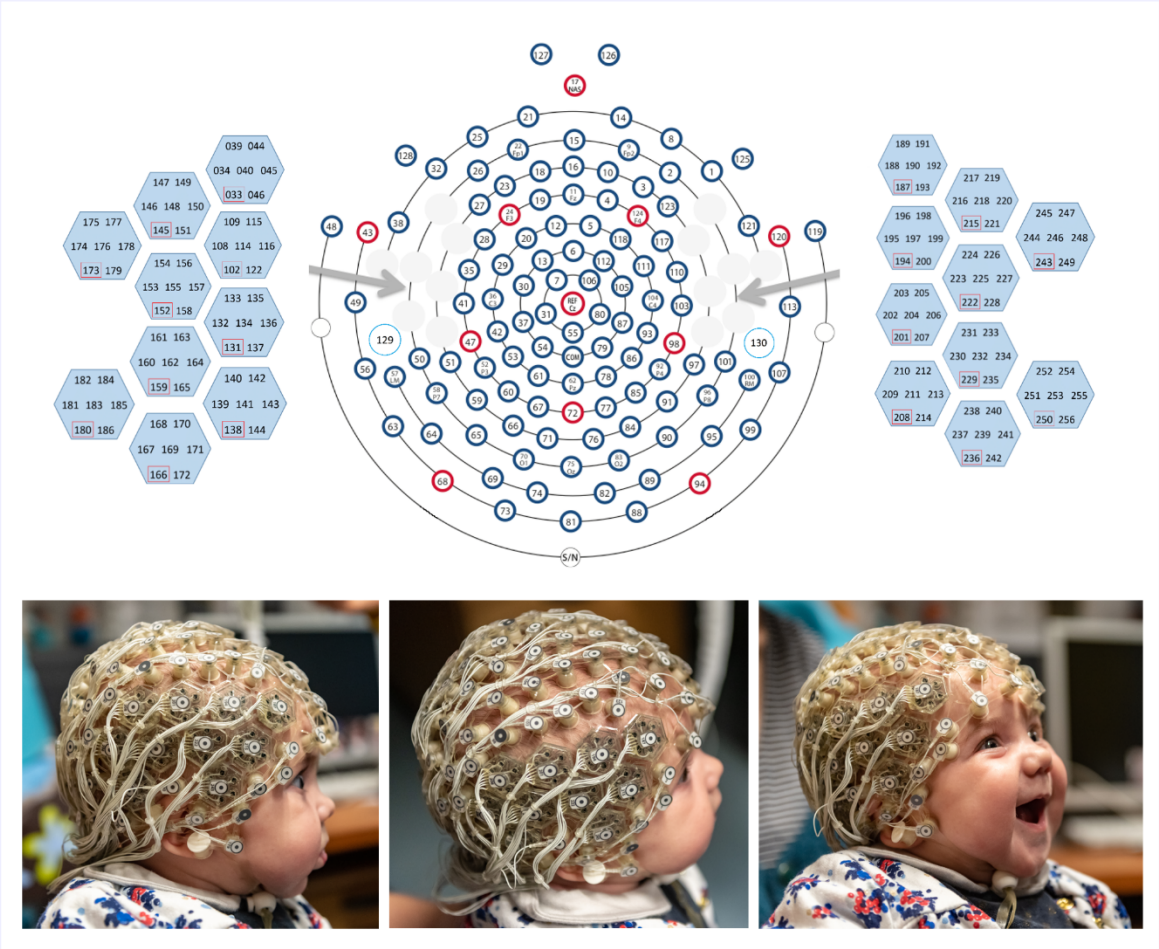

Figure 1.3 Prototype ultra-high density EEG net. Starting from the classical geodesic 128-locations partitioning (Tucker, 1993), twenty of the standard temporal positions are filled with hexagonal pods, each composed of 7 sensors with no sponge inserts. The resulting grids include 140 electrodes (70/side) displaced at a reciprocal distance of 5 mm. Sensors are made of carbon fibers embedded within a plastic (ABS) substrate and coated with silver-chloride.
We are grateful to Don Tucker and Amy Rowland for their major contribution in design and manufacture. Pictures were taken and edited by Vanna Santoro. A written permission to use the identifiable images was obtained from the parents of the infant.

1.8. Bibliography

Antell, S. E., & Keating, D. P. (1983). Perception of Numerical Invariance in Neonates. Child Development, 54(3), 695–701. https://doi.org/10.2307/1130057

Aslin, R. N. (2007). What's in a look? Developmental Science, 10(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Bechtel, W. (1998). Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge in Cognitive Science. Cognitive Science, 22(3), 295–318. https://doi.org/10.1207/s15516709cog2203_2

Bell, M. A., & Wolfe, C. D. (2008). The use of the electroencephalogram in research on cognitive development. In *Developmental psychophysiology: Theory, systems, and methods* (pp. 150–170). Cambridge University Press.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F. (2008). Hearing Faces: How the Infant Brain Matches the Face It Sees with the Speech It Hears. *Journal of Cognitive Neuroscience*, *21*(5), 905–921. https://doi.org/10.1162/jocn.2009.21076

Cantrell, L., & Smith, L. B. (2013). Open questions and a proposal: A critical review of the evidence on infant numerical abilities. *Cognition*, *128*(3), 331–352. https://doi.org/10.1016/j.cognition.2013.04.008

Carey, S., & Spelke, E. S. (1994). Domain-specific knowledge and conceptual change. In *Mapping the mind: Domain specificity in cognition and culture* (pp. 169–200).

Cheour, M., Alho, K., Čeponiené, R., Reinikainen, K., Sainio, K., Pohjavuori, M., Aaltonen, O., & Näätänen, R. (1998). Maturation of mismatch negativity in infants. *International Journal of Psychophysiology*, *29*(2), 217–226. https://doi.org/10.1016/S0167-8760(98)00017-8

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., & Näätänen, R. (1998). Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience*, *1*(5), 351–353. https://doi.org/10.1038/1561

Cheour, M., H.T. Leppänen, P., & Kraus, N. (2000). Mismatch negativity (MMN) as a tool for investigating auditory discrimination and sensory memory in infants and children. *Clinical Neurophysiology*, *111*(1), 4–16. https://doi.org/10.1016/S1388-2457(99)00191-1

Cichy, R. M., Ramirez, F. M., & Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *NeuroImage*, *121*, 193–204. https://doi.org/10.1016/j.neuroimage.2015.07.011

Coch, D., & Gullick, M. M. (2012). Event-Related Potentials and Development. In E. S. Kappenman & S. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components*. https://doi.org/10.1093/oxfordhb/9780195374148.013.0235

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, *92*(2), 225–234. https://doi.org/10.1016/j.nlm.2008.06.002

Courchesne, E. (1990). Chronology of postnatal human brain development: Event-related potential, positron emission tomography, myelinogenesis, and synaptogenesis studies. In *Event-related brain potentials: Basic issues and applications* (pp. 210–241). Oxford University Press.

deCharms, R. C., & Zador, A. (2000). Neural Representation and the Cortical Code. *Annual Review of Neuroscience*, *23*(1), 613–647. https://doi.org/10.1146/annurev.neuro.23.1.613

Dehaene-Lambertz, G. (2002). Functional Neuroimaging of Speech Perception in Infants. *Science*, *298*(5600), 2013–2015. https://doi.org/10.1126/science.1077066

Dehaene-Lambertz, G., & Gliga, T. (2004). Common neural basis for phoneme processing in infants and adults. *Journal of Cognitive Neuroscience*, *16*(8), 1375–1387.

Dehaene-Lambertz, G., & Spelke, E. S. (2015). The Infancy of the Human Brain. *Neuron*, *88*(1), 93–109. https://doi.org/10.1016/j.neuron.2015.09.026

Eggermont, J. J., & Moore, J. K. (2012). Morphological and Functional Development of the Auditory Nervous System. In L. Werner, R. R. Fay, & A. N. Popper (Eds.), *Human Auditory Development* (pp. 61–105). Springer. https://doi.org/10.1007/978-1-4614-1421-6_3

Fodor, J. A. (1983). Representations: Philosophical Essays on the Foundations of Cognitive Science. MIT Press.

Fujioka, T., Mourad, N., He, C., & Trainor, L. J. (2011). Comparison of artifact correction methods for infant EEG applied to extraction of event-related potential signals. *Clinical Neurophysiology*, *122*(1), 43–51. https://doi.org/10.1016/j.clinph.2010.04.036

Gopnik, A., & Meltzoff, A. N. (1998). *Words, Thoughts, and Theories*. MIT Press.

Grieve, P. G., Emerson, R. G., Isler, J. R., & Stark, R. I. (2004). Quantitative analysis of spatial sampling error in the infant and adult electroencephalogram. *NeuroImage*, *21*(4), 1260–1274. https://doi.org/10.1016/j.neuroimage.2003.11.028

Hamamouche, K., & Cordes, S. (2019). Number, time, and space are not singularly represented: Evidence against a common magnitude system beyond early childhood. *Psychonomic Bulletin & Review*, *26*(3), 833–854. https://doi.org/10.3758/s13423-018-1561-3

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Haxby, J. V. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. 25.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

He, C., & Trainor, L. J. (2009). Finding the Pitch of the Missing Fundamental in Infants. *Journal of Neuroscience*, *29*(24), 7718–8822. https://doi.org/10.1523/JNEUROSCI.0157-09.2009

Horst, S. (2003). The Computational Theory of Mind. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information Stanford University. https://stanford.library.usyd.edu.au/archives/sum2015/entries/computational-mind/

Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct Cerebral Pathways for Object Identity and Number in Human Infants. *PLOS Biology*, *6*(2), e11. https://doi.org/10.1371/journal.pbio.0060011

Jacobs, A. L., Fridman, G., Douglas, R. M., Alam, N. M., Latham, P. E., Prusky, G. T., & Nirenberg, S. (2009). Ruling out and ruling in neural codes. *Proceedings of the National Academy of Sciences*, *106*(14), 5936–5941. https://doi.org/10.1073/pnas.0900573106

Jusczyk, P. W. (2000). *The discovery of spoken language* (1st MIT Press pbk. ed). MIT Press.

Kabdebon, C., Leroy, F., Simmonet, H., Perrot, M., Dubois, J., & Dehaene-Lambertz, G. (2014). Anatomical correlations of the international 10–20 sensor placement system in infants. *NeuroImage*, *99*, 342–356. https://doi.org/10.1016/j.neuroimage.2014.05.046

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage*, *56*(2), 411–421. https://doi.org/10.1016/j.neuroimage.2011.01.061

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*(4), 649–662. https://doi.org/10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. *Annual Review of Neuroscience*, *42*(1), 407–432. https://doi.org/10.1146/annurev-neuro-080317-061906

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kujala, A., Huotilainen, M., Hotakainen, M., Lennes, M., Parkkonen, L., Fellman, V., & Näätänen, R. (2004). Speech-sound discrimination in neonates as measured with MEG. *NeuroReport*, *15*(13), 2089–2092.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16000960

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. https://doi.org/10.1016/j.neubiorev.2009.07.008

Luck, S. J. (2014). An introduction to the event-related potential technique (Second edition). The MIT Press.

Luhmann, H. J., Kirischuk, S., & Kilb, W. (2018). The Superior Function of the Subplate in Early Neocortical Development. *Frontiers in Neuroanatomy*, *12*, 97. https://doi.org/10.3389/fnana.2018.00097

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

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). Multiple cues for quantification in infancy: Is number one of them? *Psychological Bulletin*, *128*(2), 278–294. https://doi.org/10.1037/0033-2909.128.2.278

Molnár, Z., Luhmann, H. J., & Kanold, P. O. (2020). Transient cortical circuits match spontaneous and sensory-driven activity during development. *Science*, *370*(6514), eabb2153. https://doi.org/10.1126/science.abb2153

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109. https://doi.org/10.1093/scan/nsn044

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313–329. https://doi.org/10.1016/0001-6918(78)90006-9

Naik, S., Adibpour, P., Dubois, J., Dehaene-Lambertz, G., & Battaglia, D. (2021). *Structured Modulations of Ongoing Variability by Task and Development* (p. 2021.03.07.434162). https://www.biorxiv.org/content/10.1101/2021.03.07.434162v1

Oakes, L. M. (2010). Using Habituation of Looking Time to Assess Mental Processes in Infancy. *Journal of Cognition and Development*, *11*(3), 255–268. https://doi.org/10.1080/15248371003699977

Odabaee, M., Freeman, W. J., Colditz, P. B., Ramon, C., & Vanhatalo, S. (2013). Spatial patterning of the neonatal EEG suggests a need for a high number of electrodes. *NeuroImage*, *68*, 229–235. https://doi.org/10.1016/j.neuroimage.2012.11.062

Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: Combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, *19*(3), 162–172. https://doi.org/10.1016/j.tics.2015.01.002

Perani, D., Saccuman, M. C., Scifo, P., Spada, D., Andreolli, G., Rovelli, R., Baldoli, C., & Koelsch, S. (2010). Functional specializations for music processing in the human newborn brain. *Proceedings of the National Academy of Sciences*, *107*(10), 4758–4763. https://doi.org/10.1073/pnas.0909074107

Picton, T. W., & Taylor, M. J. (2007). Electrophysiological Evaluation of Human Brain Development. *Developmental Neuropsychology*, *31*(3), 249–278. https://doi.org/10.1080/87565640701228732

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*(2), 125–132. https://doi.org/10.1038/35039062

Pujol, J., Soriano-Mas, C., Ortiz, H., Sebastian-Galles, N., Losilla, J. M., & Deus, J. (2006). Myelination of language-related areas in the developing brain. *Neurology*, *66*(3), 339–343. https://doi.org/10.1212/01.wnl.0000201049.66073.8d

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *The British Journal for the Philosophy of Science*, *70*(2), 581–607. https://doi.org/10.1093/bjps/axx023

Ross-sheehy, S., Oakes, L. M., & Luck, S. J. (2003). The Development of Visual Short-Term Memory Capacity in Infants. *Child Development*, *74*(6), 1807–1822. https://doi.org/10.1046/j.1467-8624.2003.00639.x

Schöner, G., & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review*, *113*(2), 273–299. https://doi.org/10.1037/0033-295X.113.2.273

Slater, A., Mattock, A., & Brown, E. (1990). Size constancy at birth: Newborn infants' responses to retinal and real size. *Journal of Experimental Child Psychology*, *49*(2), 314–322. https://doi.org/10.1016/0022-0965(90)90061-C

Slater, A., & Morison, V. (1985). Shape Constancy and Slant Perception at Birth. *Perception*. https://doi.org/10.1068/p140337

Sokolov, E. N. (1963). Perception and the conditioned reflex.

Stanley, G. B. (2013). Reading and writing the neural code. *Nature Neuroscience*, *16*(3), 259–263. https://doi.org/10.1038/nn.3330

Stokes, M. G., Wolff, M. J., & Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. *Trends in Cognitive Sciences*, *19*(11), 636–638. https://doi.org/10.1016/j.tics.2015.08.016

Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*. https://www.science.org/doi/abs/10.1126/science.290.5500.2319

Thomas, D. G., & Crow, C. D. (1994). Development of evoked electrical brain activity in infancy. In *Human behavior and the developing brain* (pp. 207–231). The Guilford Press.

Tucker, D. M. (1993). Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalography and Clinical Neurophysiology*, *87*(3), 154–163. https://doi.org/10.1016/0013-4694(93)90121-B

Vilain, A., Dole, M., Lœvenbruck, H., Pascalis, O., & Schwartz, J.-L. (2019). The role of production abilities in the perception of consonant category in infants. *Developmental Science*, *22*(6), e12830. https://doi.org/10.1111/desc.12830

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Werner, L., Fay, R. R., & Popper, A. N. (Eds.). (2012). *Human Auditory Development* (Vol. 42). Springer New York. https://doi.org/10.1007/978-1-4614-1421-6

Wess, J. M., Isaiah, A., Watkins, P. V., & Kanold, P. O. (2017). Subplate neurons are the first cortical neurons to respond to sensory stimuli. *Proceedings of the National Academy of Sciences*, *114*(47), 12602–12607. https://doi.org/10.1073/pnas.1710793114

Winkler, I. (2007). Interpreting the Mismatch Negativity. *Journal of Psychophysiology*, *21*(3–4), 147–163. https://doi.org/10.1027/0269-8803.21.34.147

Yakovlev, P., & Lecours, A. R. (1967). The myelogenetic cycles of regional maturation of the brain. In *Regional development of the brain in early life* (pp. 3–70). Blackwell.

Chapter 2. ORTHOGONAL NEURAL CODES FOR SPEECH IN INFANTS

ABSTRACT

Creating invariant representations from an ever-changing speech signal is a major challenge for the human brain. Such an ability is particularly crucial for preverbal infants who must discover the phonological, lexical and syntactic regularities of an extremely inconsistent signal in order to acquire language. Within the visual domain, an efficient neural solution to overcome variability consists in factorizing the input into a reduced set of orthogonal components. Here, we asked whether a similar decomposition strategy is used in early speech perception. Using a 256-channel electro-encephalographic (EEG) system, we recorded the neural responses of 3-month-old infants to 120 natural consonant-vowel syllables with varying acoustic and phonetic profiles. Using multivariate pattern analyses, we show that syllables are factorized into distinct and orthogonal neural codes for consonants and vowels. Concerning consonants, we further demonstrate the existence of two stages of processing. A first phase is characterized by orthogonal and context-invariant neural codes for the dimensions of manner and place of articulation. Within the second stage, manner and place codes are integrated to recover the identity of the phoneme. We conclude that, despite the paucity of articulatory motor plans and speech production skills, pre-babbling infants are already equipped with a structured combinatorial code for speech analysis, which might account for the rapid pace of language acquisition during the first year.

2.1. INTRODUCTION

A major, fundamental challenge for any brain is to build stable representations of a changing world. In particular regarding speech, the breadth of the human lexicon and its possibilities of morphemic composition are based on fine phonetic differences that undergo substantial acoustic restructuring depending on many contextual factors such as voice peculiarities, intonation and co-articulation. Nonetheless, we effortlessly perceive "bog" and "dog" as steady and distinct words, no matter whether shouted by a little girl or whispered by an elderly man. The capacity to extract invariant neural representations from the extremely variable speech signal is essential for adults and even more crucial for infants, who must discover the organizing regularities of speech in order to acquire their native language. Yet, the neural underpinnings of such an ability remain underspecified.

In the visual domain, recent findings, based on neuronal recordings during object (Behrens et al., 2018) and face recognition (L. Chang & Tsao, 2017), suggest that in order to deal with the large amount of incoming pictures, the brain factorizes the input into independent and orthogonal low-dimensional components, each coding for a different dimension of variation. For instance, faces may be decomposed into as little as 50 orthogonal dimensions, thus effecting a remarkable dimensional reduction (L. Chang & Tsao, 2017). The components are thought to be subsequently recombined to yield unified percepts. Can such an account be applied to speech? Apart from any neural consideration, linguists have defined phonemes as bundles of a small set of orthogonal phonetic features, each corresponding to a binary code that summarizes an articulatory dimension and its acoustic correlates (Halle, 2013). For instance, the phonemes "b" and "d" from the example above share all parameters (+consonantal and -vocalic, +obstruent and -sonorant, +voiced, etc.) except for the place of articulation (+labial/-alveolar vs. +alveolar/-labial). Given their linguistic characteristics (distinctive, minimal and combinable), these features might correspond to the basic decomposition axes harnessed by the brain to reduce the high dimensionality of the input, thereby overcoming speech variability.
In the last years, high-resolution intracranial recordings on adults (Mesgarani et al., 2014) and fMRI adult data (Arsenault & Buchsbaum, 2015; Correia et al., 2015) have provided evidence in line with this hypothesis: a partial neural specialization for phonetic features was observed during passive listening of speech. Here, we ask whether such a decomposition strategy is already present in early infancy.

The first essential step for language acquisition consists in the identification of the native sound structure. Delineating the type of speech representations infants start with is thus crucial to elucidate how they can discover the phonetic repertoire and phonological grammar of their native tongue. A plethora of classical studies has demonstrated that infants come to the world with the perceptual abilities necessary to distinguish a variety of phonetic contrasts (Bertoncini et al., 1987; Eimas et al., 1971; Eimas & Miller, 1980a among others). Moreover, both behavioral and neuroimaging

researches have shown that, since birth, they spontaneously override the acoustic variability produced by changes in talker's voice (Dehaene-Lambertz & Pena, 2001; Jusczyl et al., 1992), speaking rate (Eimas & Miller, 1980b; Miller & Eimas, 1983) and prosody (Fló et al., 2019). Interestingly, the type of perceptual constancy newborns exhibit corresponds precisely to that required to establish reliable links between speech sound differences and changes in meaning. Although remarkable, the early ability to detect minimal phonetic contrasts among syllables does not truly inform upon the nature of the underlying neural code: infants might either process utterances as integral wholes (e.g. in the form of broad spectro-temporal patterns organized around sonorous nuclei) or decompose them into smaller elements (e.g. phonemes or phonetic features).

Behavioral investigations have shown that newborns and 2-month-olds fail at identifying a shared consonant in a group of syllables containing different vowels (Bertoncini et al., 1988; Jusczyk & Derrah, 1987). Further, neonates proved capable of categorizing utterances using the number of their syllabic constituents but not the number of phonemes (Bijeljac-Babic et al., 1993). Following these results, many authors have proposed the syllable as the primitive unit for speech processing. Computational modelling has corroborated the plausibility of this conclusion by showing that sonority-based syllable-like structures are indeed accessible, in conversational speech, by means of general auditory mechanisms (Räsänen et al., 2018). Currently, such kind of broad and holistic units is widely assumed to be the starting point for lexical learning when no linguistic knowledge is available.

However, progress in neuroimaging has opened the way to new paradigms that, bypassing behavioral limitations, may uncover the existence of unexpectedly refined abilities early in development. Following the repetition of CV (consonant-vowel) syllables differing only in their vocalic component, EEG recordings revealed that 3-month-olds could recognize the shared consonant and detect when it changed (Mersad & Dehaene-Lambertz, 2016). They could even learn to associate each consonant to a visual shape, independently of the vocalic surroundings (Mersad et al., 2021). Such finding, easily explicable in terms of sub-syllabic processing, prompts to re-examine the format of early speech representations.

To this aim, we combined high-resolution EEG recordings with time-resolved multivariate pattern analysis. Twenty-five 3-month-old infants were exposed to 120 natural consonant-vowel syllables, presented in pseudo-random order during about one hour. Syllables were chosen to independently vary the consonantal dimensions of manner (obstruent vs. sonorant) and place of articulation (labial vs. alveolar vs. velar). Each consonant was coupled with two vowels (/i/ and /o/) and produced by a male and a female speaker in five distinct utterances, to ensure acoustic and co-articulatory variability across tokens with the same phonetic profile (Figure 2.1A). The dimensions of manner and place of articulation were chosen due to the highly contrasted levels of consistency characterizing their acoustic correlates: whereas manners are reflected in prominent spectrotemporal prototypes (Stevens, 2000), the acoustic cues for place are more subtle (Shannon et al., 1995) and complex (Smits et al., 1996), hence fundamentally dependent on the context of

production (Fowler, 1994). Such acoustical divergence was especially evident in the auditory similarity structure of our stimuli set, as illustrated in section 2.5.1 and Figure 2.5.

We used multivariate decoding analyses to investigate infant speech processing at three possible levels corresponding to holistic syllables, phonemes and phonetic features[12]. Linear classification algorithms are powerful tools in that they can combine multiple sources (here EEG channels) to find the optimal combination of brain signals reflecting the variables of interest (Hebart & Baker, 2018). Since *any* peculiarity in the data can be used to separate classes, showing that neural responses can be sorted according to certain labels, in itself, does not speak to the underlying encoding scheme. A key strategy in this regard consists in examining the pattern of generalization: how decoders trained in a particular context perform across variations that are expected to be non-pertinent for a given code (Kriegeskorte & Douglas, 2019). For instance, if infants extract speaker-invariant information, then decoders trained on the brain responses to syllables produced by the male voice are expected to generalize to the female voice (and vice-versa). This logic was central to the purpose of the present study. We reasoned that, if consonants and vowels were processed separately, then a decoder trained in the context of, say, vowel "o", should generalize to the context of the other vowel "i". Conversely, such generalization should not be possible if each syllable was encoded by its own idiosyncratic neural code. At the sub-syllabic level, we could ask if a decoder trained to separate "bo" vs "do" is able to 1) correctly classify "mi" vs "ni", thus revealing the presence a neural code for the places "labial" vs "alveolar" that is orthogonal to vowels and manners; or 2) generalize only to "bi" vs "di", thus indicating an idiosyncratic and integrated neural code for the consonants "b" vs "d", without further decomposition into separable dimensions.

Furthermore, using time-resolved EEG signals, it is possible to train a distinct decoder at each time point to probe the presence of distinct patterns of generalization over time (King & Dehaene, 2014). By tracing the time-course of generalizations and class confusability, we could ask whether and when particular pieces of information were re-coded across stages of processing. A factorized encoding model, similar to that observed for faces (L. Chang & Tsao, 2017), predicts an early projection of the signal into a small set of orthogonal dimensions, followed by their integration into broader chunks (consonants/vowels or even entire syllables). The opposite decomposition process, progressing from holistic syllables to phonemes or/and features, is also imaginable.

Decoding speech from noisy infant event-related potentials (ERPs) is a difficult task. To enable it, we recorded a large data set consisting of ~3100 trials/participant. Furthermore, we collected ERPs with a high-density custom net featuring an unusual number of 256 channels (Figures 2.1B and Figure 1.3; see also Figure 2.1C for the grand average across all syllables). This novel intensive electrode coverage, combined with the thinness of infant skulls, should enhance the spatial resolution of our

[12]for the moment, the terms "syllable", "phoneme" and "phonetic feature" are used as convenient stimuli descriptors, regardless of the acoustic/linguistic value they might hold for the brain.

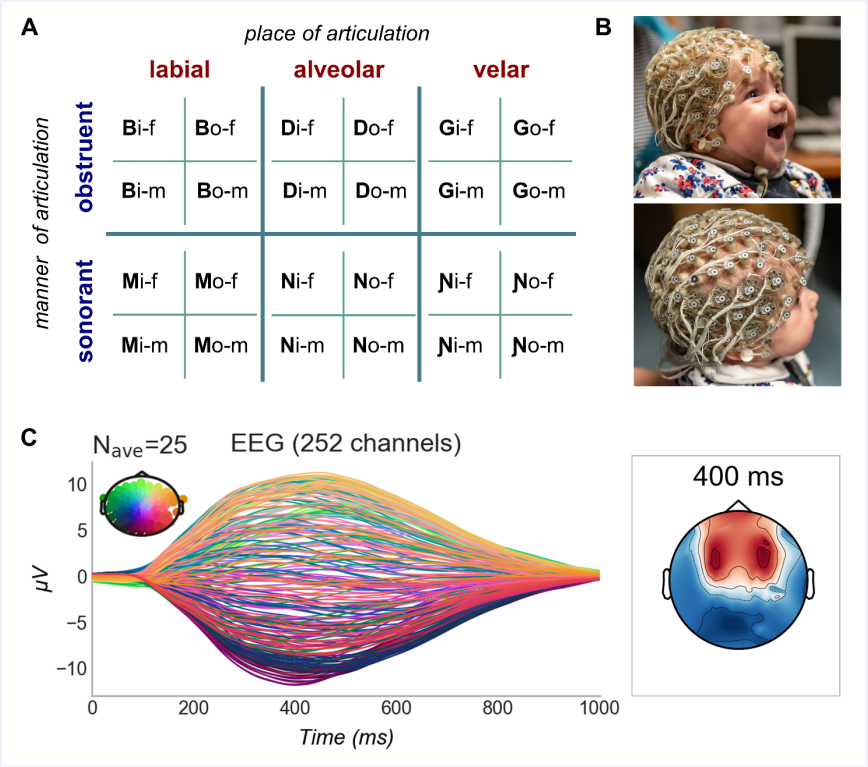

Figure 2.1 Experimental set-up and average syllable-related potential.
(A) Stimuli sub-conditions and their phonetic characteristics (f=female, m=male voice). (B) 256 channels super-high-density net on a 3-month-old infant: tight grids of custom electrodes are arranged over the auditory linguistic areas of the superior temporal lobe. (C) Grand average ERP: all conditions are pulled together.

recordings and facilitate the discrimination of ERPs arising from spatially close neuronal clusters (Stokes et al., 2015).

2.2. MATERIAL & METHODS

2.2.1. Participants

25 full-term, normal-hearing infants (12 females, 13 males) coming from a French-speaking environment were tested between 12 and 14 weeks after birth (mean age= 12 weeks and 6 days). An additional 16 participants were excluded from analysis because of: excessive agitation during the experimental session (n=6), insufficient number of trials after artifact rejection (n=3, the artifact rejection procedure is described below), technical problems during data collection (n=3), aberrant global field power (GFP) in the average of all syllable-related potentials (i.e. peak GFP<4μV, n=4). The protocol was approved by the regional ethical committee for biomedical research (CPP Region Centre Ouest 1). Parents gave their written informed consent before starting the experiment.

2.2.2. Stimuli

Stimuli consisted of 120 speech sounds constructed upon 6 consonants: /b/, /d/, /g/, /m/, /n/, /ŋ/. These consonants were selected to cover two manner features, i.e. obstruent (/b/, /d/, /g/) and sonorant (/m/, /n/, /ŋ/), and three places of articulation, i.e. labial (/b/, /m/), alveolar (/d/, /n/), and palatal-velar (/g/, /ŋ/, referred to as "velar" for simplicity). Each consonant was associated with

two vowels, /i/ and /o/, and produced by a male and female speaker to obtain 2 manner *x* 3 place *x* 2 vowel *x* 2 voice factor design (i.e. 24 sub-conditions). To increase acoustic variability (and extend the external validity of our measurements), speakers were asked to repeat the same tokens several times changing their intonation. For every sub-condition we selected 5 utterances, distinct in low-level acoustic characteristics such as pitch and duration. In the resulting set of syllables each manner of articulation condition contained 60 spectrotemporal profiles (3 consonants *x* 2 vowels *x* 2 voices *x* 5 utterances); similarly, each place of articulation was presented in 40 (2 consonants *x* 2 vowels *x* 2 voices *x* 5 utterances) spectrotemporal versions.

Stimuli construction

Speech signals were recorded in a silent chamber using a dynamic microphone (Beyerdynamic DT 290 broadcast headset) on a linear PCM recorder (DR-05, TASCAM) at a sampling rate of 44.1 kHz. Recordings were first cleared from background noise in Audacity 2.1.3 (https://www.audacityteam.org) and further edited with PRAAT software (Boersma & Weenink, 2017). Acoustic transients (clicks) were manually removed and stimuli length was adjusted to fall within the range of 350-425ms. Tokens were normalized for peak amplitude and average (i.e. root-mean square) intensity, obtaining maximal audibility and loudness equalization. All stimuli were placed on the left channel and a click was positioned on the right channel at the exact time-point of syllable onset. The left channel was connected to the audio amplifier (mono input to the loudspeakers) while the right channel was connected to the EEG amplifiers through the DIN port to create a TTL signal. Brain voltage and clicks were recorded simultaneously with the same temporal resolution providing a precise mapping between EEG recording and stimulation.

Articulation, and in particular the manner, is known to affect consonant duration, introducing the risk of possible confounds between this low-level cue and the phonetic feature. To validate our set of syllabic stimuli, we therefore assessed consonant lengths through a gating procedure (Grosjean, 1996). Over multiple trials, each stimulus was listened in portions of progressively increasing duration (10ms steps), starting from the end of the syllable and proceeding backwards, toward its beginning. The duration of the longest portion for which no consonantal sound was perceived was subtracted from the total length of the stimulus. Consonant duration assessed in this way ranged between 80 and 210 ms (M±SD=154±25) and varied homogeneously across categories (i.e. /b/, /d/, /g/, /m/, /n/, /ɲ/; $F(5,114)=1.42$, $p=0.222$). Most importantly, consonant duration did not change as a function of manner nor place of articulation. In an ANOVA with these two factors, the effect of manner ($F(1,114)<1$), the effect of place ($F(2,114)=1.28$, $p=0.280$) and their interaction ($F(2,114)=2.25$, $p=0.109$) were not significant.

2.2.3. Procedure

Subjects were tested in a soundproof Faraday cage equipped with a computer screen and loudspeakers on the top. Infants were hold by a caregiver, their position was chosen to guarantee

personal comfort and at the same time enable good-quality data acquisition. Syllables were broadcast through the loudspeakers at 70 decibels, in a latin-square randomized order and with a randomly selected inter-stimulus interval (ISI) between 600 and 1000ms. To minimize body movements we presented engaging visual animations that were unsynchronized with the auditory stream. Sleep was highly encouraged at any time; on average our subjects slept for 65% of the experimental session. Breaks were taken whenever necessary. The experiment finished with the presentations of 3136 tokens (corresponding to approximately 63 minutes of listening time) or as soon as infants became restless.

2.2.4. EEG recording and data preprocessing

The electroencephalogram (EEG) was continuously digitized at 500 Hz (Net Amps 300 EGI amplifier combined with NetStation 5.3 software) from 256 channels. We used a prototype HydroCel net (EGI; Eugene, OR, USA) referenced to the vertex. In this customized net the surface corresponding to twenty temporal locations in the classical geodesic layout (128 partitioning) is filled by 2 tight grids of sensors (70 on each side of the head) displaced at a reciprocal distance of 5 mm (Figure 1.3).

Artifact detection and correction

Data preprocessing was conducted through custom-made MATLAB scripts based on the EEGLAB toolbox 14.0 (Delorme & Makeig, 2004). While following the main preprocessing steps normally used in developmental studies, we introduced some modifications inspired by efforts carried to improve adult data quality (Jas et al., 2017; Mognon et al., 2011). Namely, we identified artifacts on the continuous recordings with the employment of adaptive rather than absolute/predefined thresholds. In this way, we could account for inter-individual variability and the heterogeneous influence that reference distance and vigilance state exert on the voltage. Moreover, we did not discard but corrected local and transient artifacts, exploiting the redundancy of information provided by our dense sensor-layout (Figures 1.3 and 2.1B) and high sampling rate.

As a first step, EEG recordings were band-pass filtered ([0.5 - 40Hz]) and the mean voltage of each electrode was set to zero. Artifacts were detected before segmentation by a series of algorithms with adaptive thresholds. These algorithms rejected samples on the basis of: the voltage amplitude and its first derivative; the variance across a 500ms-long moving time window; the fast running average and the deviation between the fast and the slow running averages within a 500ms-long sliding time window. Thresholds were set independently for each subject and for each electrode upon the distribution of these measures along the whole recording (threshold = median +/- n*IQ, where IQ is the interquartile range of the distribution). Two additional algorithms identified whether the power within the 0-10Hz band was excessively low or within 20-40Hz excessively high relative to the total power; and whether the voltage amplitude displayed by each sensor at a given time point was disproportionate relative to that recorded by the other sensors at the same instant. For these last two algorithms, thresholds were computed upon the distribution across channels.

The output of the artifact detection procedure was a rejection matrix with the same size of the EEG recording. We used this matrix to mark time points with prominent artifacts (*bad times*) and channels that did not function properly (*bad channels*). We identified as *bad times* periods longer than 50ms with a percentage of rejected channels superior to 30% or beyond 2IQ from the 3rd quartile of the distribution of the percentage of rejected channels across time. Similarly, *bad channels* were the ones not working properly for more than 30% of time or with a percentage of bad samples that went beyond 2IQ from the 3rd quartile of the distribution of the percentage of rejected samples across channels.

Periods defined as *bad times* were not corrected because there was not enough information available to reconstruct the signal. For the rest, two kinds of correction were applied. When the rejected segments had a very short duration (50ms max, e.g. heart beats or jumps) we relied on the assumption that, during these periods, most of the variance came from noise. For each of them, principal components were estimated (PCA) and the first *n* components determining 90% of the variance were removed. Otherwise, we corrected *bad channels* and long rejected segments that did not contain *bad times* using spherical splines interpolation (Perrin et al., 1989). Spatial interpolation was carried out only if at least 50% of the neighboring channels were intact. Corrected segments were realigned with the rest of the data which were then high-pass filtered (0.5Hz) to eliminate possible drifts resulting from this operation.
The artifact detection-correction procedure was applied iteratively, keeping previously identified bad samples aside for the subsequent artifact detection steps.

Epoching

EEG recordings (and the corresponding rejection matrix) were segmented into epochs starting 200ms before and ending 1400ms after syllable onset. Trials were rejected if more than 15% of their samples contained artifacts. Epochs were also discarded based on their Euclidean distance from the average, i.e. when their mean or maximum distance from the average response was an outlier in the distribution (> 3rdquartile + 1.5*IQ). Following automated rejection, the remaining epochs were visually inspected and a few trials still presenting obvious aberrancies were manually eliminated.

Since multivariate pattern analysis requires a conspicuous amount of trials, we included subjects with a minimum of 40 trials/sub-condition. In our final group of infants (N=25), the mean trial rejection rate was 28.7% (12.4 to 53.5%). On average, the number of artifact-free epochs available per subject in each sub-condition (e.g. “bi-female”) was 70, providing 840 trials for each manner of articulation condition and 560 trials for every place of articulation condition.

Before submitting them to the main analyses, epochs were low-pass filtered at 20Hz, mathematically re-referenced to the mean of all channels and down-sampled (with a moving average of 2 time points) to 250Hz. All the main analyses (decoding) were carried at the single trial level. Nonetheless, epochs

were also averaged per either sub-condition or manner-/place-condition in order to examine evoked responses (ERPs, e.g. Figure 2.7C).

2.2.5. Decoding

Multivariate pattern analyses were conducted within subject, relying on the Scikit-Learn (Pedregosa et al., 2011) and MNE (Gramfort et al., 2013, 2014) Python packages. To decode *in time* epochs were divided into 60 consecutive windows of 20ms (from -200ms to 1000ms relative to stimulus onset), each corresponding to a matrix with the shape *n* channels *x* 5 samples (sampling rate = 250Hz, 5 samples=20ms). Each analysis was carried on a single window with the general aim of predicting a vector of categorical data (*y*) from a matrix of single-trial neural data (*X)* which included all EEG channels. To decode the manner of articulation trials were labelled as belonging to either the category of "obstruent" or to the category of "sonorant" depending on whether /b/, /d/, /g/ or /m/, /n/, /ɲ/ exemplars were presented. To decode the place of articulation *y* comprised three classes: "labial" (/b/, /m/), "alveolar" (/d/, /n/), and "velar" (/g/ and /ɲ/). For vowel decoding, trials were separated in two classes, "i" and "o", based on the vocalic portion of the stimulus.

All decoding analyses were performed within a stratified cross-validation procedure consisting of 100 iterations. At each run, trials were shuffled and then split into a training and a test set containing 90% and 10% of trials respectively. As compared to the most common folding approach, this cross-validation outline enabled to maximize the number of iterations (and thus the reliability of the final performance) while maintaining a fixed and reasonable amount of test trials. Importantly, stratification ensured (a) that the same proportion of each class was preserved within each set (b) all sources of variability (e.g. voice gender) were evenly represented across sets (e.g. training and test sets contained syllables produced by the female vs male speaker in the same proportion).

Given the high amplitude fluctuations typically seen in infant EEG background activity, we first aimed at improving our signal-to-noise ratio. Once defined the training and the test set for a given run, we applied a "micro-averaging" procedure, a strategy previously used on adults with the same purpose (Grootswagers et al., 2016). This consisted in averaging together randomly picked groups of 16 epochs within each class. The number of trials to average being arbitrary, we tried with 4, 8, and 12 and observed that by averaging 16 trials we could reach the best performance without compromising its reliability. Note that such assessment was conducted on the first decoding analysis we had planned (i.e. manner of articulation within a standard cross-validation schema) and the choice of 16 was then adopted a priori for all the other decoding analyses. At the end of this operation, to ensure perfect balance among classes, we equalized the number of (micro-averaged) epochs across categories. In practice, this consisted in dropping 1 to 3 randomly picked trials from the most numerous class(es).

Next, following the z-scoring each feature (i.e. channel and time point across trials), a L1-norm regularized Logistic Regression (Fan et al., 2008) was fitted to the training set in order to find the

hyperplane that could maximally predict *y* from *X* while minimizing a log loss function. L1 penalty was chosen to exclude less informative features from the solution (their weights being set to zero). Such regularization can be conceived in terms of dimensionality reduction, an optimization that enabled us to prevent overfitting (by reducing model complexity (Ng, 2004)) but still exploit the high density of our EEG data. The other model parameters were kept to their default values as provided by the Scikit-learn package. When decoding concerned more than two classes (e.g. place classification) we adopted a "one-vs-rest" approach: for each class (i.e. each place of articulation) one model was fitted against all the other classes.

Once trained, the models were used to predict *y* from the test set and their performance was evaluated by comparing estimates to the ground truth. All algorithms produced as an outcome vectors of probabilistic estimates. These probabilities were scored by computing the area under the Receiver Operating Characteristic curve (AUC), which summarizes the ratio between true positives (e.g. trials correctly classified as "obstruent") and false positives (e.g. trials classified as "obstruent" while a sonorant consonant was presented). The value of AUC ranges between 0 and 1, with 0.5 corresponding to chance level. Once again, in multiclass decoding a "one-vs-rest" scheme was used: the AUC scores were computed for each class against all the others and then averaged. Lastly, for both binary and multiclass problems, evaluations were averaged over all cross-validation runs.

As a proof of concept, the main decoding analyses were performed with two additional algorithms: L1-norm regularized linear Support Vector Machine (SVM; Fan et al., 2008) and Linear Discriminant Analysis (LDA). For the latter, a shrinkage estimator of the covariance matrix was used, taking into account the fact that the dimensionality of our data vectors exceeded the number of samples in each class (Ledoit & Wolf, 2003). Importantly, we restricted our alternatives to linear classifiers to make sure that the algorithms focused on explicit neural codes (Kriegeskorte, 2011). Beside slight variations in accuracy, alternative classifiers yielded very similar outcomes.

Generalization across time

Estimators trained at each time window *t* were systematically tested on (both the same and) every other possible time window *t'*, i.e. every 20ms from 200ms prior to 1000ms after syllable onset. Such procedure was performed within the cross-validation so that training set at *t* and test set at *t'* came from different groups of trials. In the resulting "temporal generalization matrices" each row corresponds to the time lag at which the estimator was trained and columns correspond to the time windows at which it was tested (King & Dehaene, 2014). The shape of the performance within these matrices provides peculiar insights upon the dynamics of the underlying brain activity. If the same neural code was found at *t* and *t'*, the classifier trained at *t* would generalize at *t'*. If, on the contrary, information was passed to another stage of processing characterized by its own coding scheme, performance at *t'* would be at chance (King & Dehaene, 2014).

Generalization across conditions

We examined the consistency of information used by classifiers in different harmonic and co-articulatory contexts by performing cross-condition decoding. To ask whether the same neural codes supported the classification of phonetic features and vowel identities across different harmonic contexts, we trained estimators on manner contrasts (/b/, /d/, /g/ vs /m/, /n/, /ɲ/); place contrasts (/b/, /m/ vs /d/, /n/ vs /g/, /ɲ/) and vowel contrasts (/i/ vs /o/) within one speaker condition (e.g. syllables pronounced by the female voice) and tested these same estimators on the other speaker condition (e.g. syllables spoken by the male voice). The procedure regarding co-articulations was analogous: we trained place and manner estimators on one vowel context and tested them on the other; we trained vowel estimators on single manners or places and assessed their performance on the alternative ones.

To test the orthogonality of manner and place encoding we trained estimators on each featural condition separately. More specifically, to reveal place-independent phonetic processing classifiers were trained on the manner comparison ("obstruent" vs "sonorant") at single place contexts (e.g. only labial sounds). These estimators were then tested both at the trained place (e.g. labials) and at the two unseen places (e.g. alveolar and velar consonants). In case manner neural codes were independent from the place of articulation, we expected classifier to perform comparably *within* the trained place and *across* unseen place contexts. Following the same rationale, we asked whether place codes are specific to manners of articulation by training classifiers to discriminate labials vs. alveolars vs. velars on one manner (e.g. only with obstruent sounds) and testing them within the same (e.g. obstruents) and at the alternative manner condition (e.g. sonorants).

Moreover, we investigated the orthogonality of consonant and vowel codes with two complementary procedures. First, we trained algorithms to distinguish each consonant based on single vocalic contexts (e.g. separation of /b/ vs /d/ vs /g/ vs /m/ vs /n/ vs /ɲ/ when they were co-articulated with /i/) and tested them within the same and across the alternative co-articulatory context (e.g. classify consonant identity among "bo", "do", "go", "mo", "no", "ɲo"; note that for this schema, as for place classification, we adopted a "one-vs-rest" approach and the percentage of correct classifications as evaluative metric). Analogously, we trained vowel classifiers on each consonantal option and assessed their performance within the trained consonant and across the five alternative ones. In case consonant and vowel were encoded separately, we expected to obtain comparable scores *within* and *across* conditions; oppositely, a degradation in performance across conditions would be indicative of interdependence between the two.

For cross-condition decoding we modified the cross-validation scheme described above so that models fitted on each training set were directly applied at all trials belonging to the untrained condition (i.e. the test set "*across"*). In this way, we capitalized on the independence of train and test sets. Concerning the splitting of single-condition datasets (i.e. the dataset *"within"*), the number of test trials was calibrated to guarantee a minimum of 2 micro-averaged trials/class at test and at the

same time maximize the amount of trials available for training. Note also that in order to ensure an adequate number of training/test samples, the micro-averaging for the last two cross-decoding schemas was reduced to groups of 8 epochs. Apart from these modifications, the decoding procedures resembled those described above.

2.2.6. Neural syllable confusion and multiple regression analysis

For this section we first built a twelve-class decoding problem by pulling together the female and male conditions and then training algorithms to separate each syllable from all the others (i.e. “bi” vs “bo” vs “di” vs “do” vs “gi” etc.). We adopted a “one-vs-rest” approach and used the same pre-processing steps described for the main analyses. Within each cross-validation loop, we stored the error matrices displayed by these classifiers at test. After averaging across runs, we obtained a series of matrices where the entry at row *i* and column *j* corresponds to the percentage of samples belonging to class *j* and labeled as *i* by the classifier (Figure 2.4C-left and Figure 2.9A-bottom). The diagonal of these confusion matrices depicts class-wise accuracy, with theoretical chance being at 8.3% (Figure 2.9A-top). Given that there is a variety of stimuli characteristics other than syllable identity which could lead to above-chance scores (up to 50%), diagonal entries alone are hardly interpretable. On the other hand, misclassification patterns (i.e. off-diagonal entries in the matrices) have the potential to reveal which dimensions of the stimuli the neural code honors or disregards. To uncover the neural representational geometry (Kriegeskorte & Kievit, 2013) captured by our algorithms and its evolution over time, we employed multiple linear regression. Specifically, we modeled each confusion matrix as a linear combination of five classification performances: those of the ideal manner, place, consonant, vowel and whole-syllable decoders (Figure 2.4C-middle and 2.9B-top). Concerning the matrix modelling manner discrimination, for example, the predicted entries for those pairs of syllables sharing the same manner correspond to 16.6%, whereas the predicted value for pairs of syllables not sharing the same manner is 0%. The five predictors were used to explain the (neural) syllable confusion observed at each time point, generating a vector of beta-weights for each of the five regressors. All matrices were z-transformed before estimating the coefficients. Significantly above-zero beta-weights assigned to a particular regressor indicate that, at a given time point, the classifier relies on the dimension reflected by that model over and beyond the remaining four variables.

2.2.7. Statistical analysis

To calculate statistics we performed second-level tests across subjects employing the MNE dedicated functions. Following the example in (King et al., 2016), we tested whether (a) time-resolved classification scores were higher than chance; (b) time-resolved classification scores within the trained context were superior to those across context; (c) whether multiple regression beta-weights were higher than zero; using one-sample cluster-based permutation t-tests (Maris & Oostenveld, 2007) which intrinsically account for multiple comparisons. The analyses considered

one-dimensional clusters in all cases apart from the generalization across time matrices (with shape training times *x* testing times) for which clusters were bi-dimensional. Univariate t-values were calculated for every score/beta-weight with the exclusion of those corresponding to the baseline period. All samples exceeding the 95th quantile were then grouped into clusters based on cardinal or diagonal adjacency. Cluster-level test statistics corresponded to the sum of t-values within each cluster. Their significance was computed by means of the Monte-Carlo method: they were compared to a null distribution of test statistics created by drawing 10000 random sign flips of the observed outcomes. A cluster was considered as significant when its p-value was below 0.05.

2.3. RESULTS

For all the analyses described below, we trained and tested series of linear estimators on brief (20ms) consecutive windows all along the time course of the ERPs. Our goal was to define the granularity of the infant coding scheme for speech: is it syllabic, phonetic or featural?

2.3.1. Successful classification is based on dynamic and discrete neural patterns

We first assessed whether decoders trained on infant brain responses could classify the EEG recordings according to the phonetic characteristics of the speech stimuli. Figure 2.2A-B show that obstruents could be distinguished from sonorants starting from 80ms after syllable onset (p_{clust} =0.0001; peak performance observed at 200ms: N=25, M=0.735±0.08, chance= 0.5), while places of articulation were reliably classified over two time windows: 220-480ms (p_{clust}=0.0001; peak at 260ms: M=0.545±0.039); and 540-720ms (p_{clust} =0.0028; peak at 640ms: M=0.534±0.042). As for what concerns vowels, the two alternatives in our design (/i/ and /o/) differ in both height and backness, precluding the isolation of phonetic sub-classes. Nonetheless, Figure 2.2C shows that vowel identity was reliably discerned in between 260 and 600ms (p_{clust} =0.0001; peak at 480ms: M=0.596±0.08, chance=0.5) and from 760ms onwards (p_{clust} =0.0001; peak at 860ms: M=0.56±0.067, chance=0.5).

To fully characterize the neural dynamics underlying such performances, the same classifiers were systematically tested on their ability to decode across time. When neural activation is maintained over time, a successful estimator, trained at a given time point, will continue to achieve above-chance scores over a broader time range (King & Dehaene, 2014). Figure 2.2D illustrates how classifiers generalized only over a limited amount of time lags, indication that the neural activity was progressing along a functional pathway. Concretely, the "cone" shape arising from the generalization matrices discloses the retrieval of evolving neural codes: the activity supporting classification was either transferring across cortical regions, transformed within the same region over time or both.

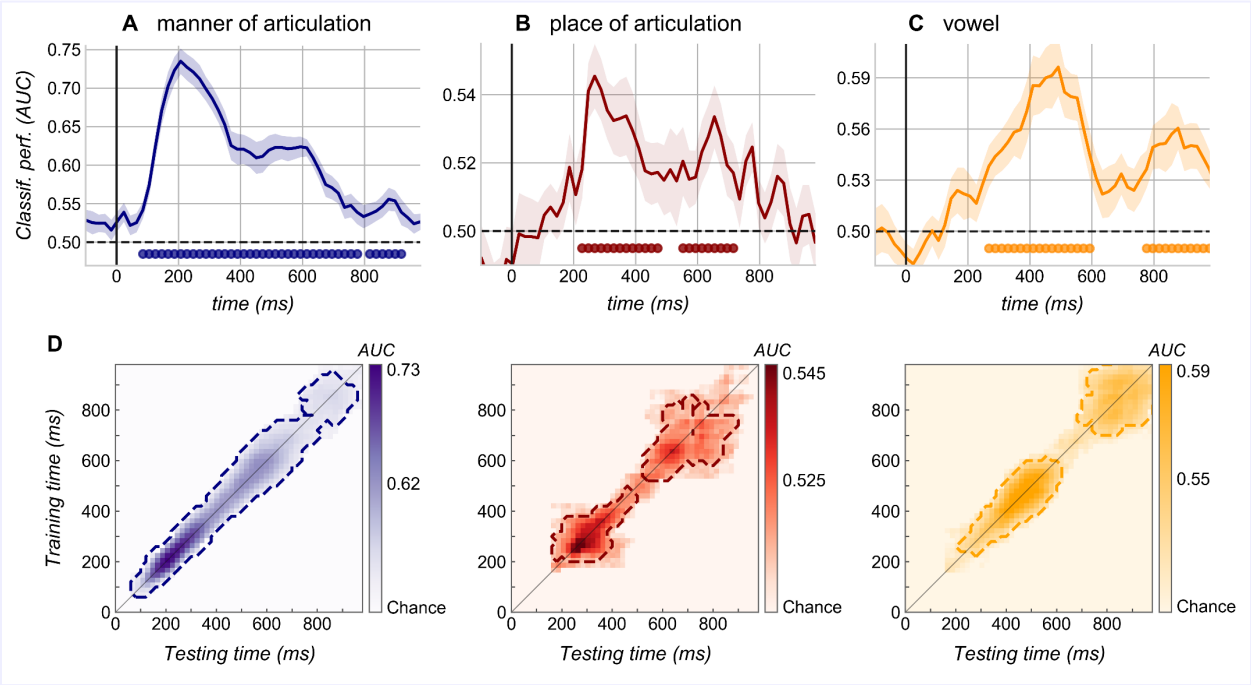

Figure 2.2 Classification performances of estimators trained on single time windows (20ms) along the ERP. Top: Estimators are tested at the trained time sample. Shaded areas correspond to the standard error (SEM) across subjects, dotted black lines mark theoretical chance level and filled circles indicate significant scores (cluster-corrected t-test). (A) Performance of classifiers trained on manner distinctions: obstruents (/b/, /d/, /g/) vs. sonorants (/m/, /n/, / ɲ/). (B) Performance of classifiers trained on place distinctions: labials (/b/, /m/) vs. alveolars (/d/, /n/) vs. velars (/g/, /ɲ/). (C) Classification of vowel identities: /i/ vs /o/. (D) Temporal generalization matrices: each panel displays above-chance decoding scores of estimators trained on a single time window (y-axis) and tested at every possible time sample (x-axis) along the ERP. The diagonal thin lines demark classifiers trained and tested on the same time sample. Dashed contours indicate significant clusters (manner: p_{clust} =0.0001; place: p_{clust} =0.0001 and 0.0028, vowel: p_{clust} =0.0097 and p_{clust} =0.0108).

Presumably, the mild widening of the generalization performance observable in the second portion of the trial denotes a change in the representational format reached relatively late after syllable onset.

To objectivize this interpretation we used classifier weights to reconstruct informative activity patterns (see section 2.5.2). Discriminative activity was diffuse over the scalp, resembling the auditory ERP topographies arising from multiple perisylvian sources that are typical of this age. Crucially, informative clusters were qualitatively different during the first and second time-windows of reliable classifiability (Figure 2.6), substantiating the occurrence of distinct encoding stages. Change was particularly appreciable in the individual topographies (Figure 2.6A-B) which are free of the blurring effect created by averaging across participants. We additionally observed that sensors supporting manner and place classification were somewhat separable (Figure 2.6); and found significant differences between brain activity patterns precisely distinctive for either labials, alveolars or velars (Figure 2.7, where a detailed overview of place-informative activations is also

reported). These findings uncover that infant syllable perception is supported by spatially distinct, although distributed and partially overlapping, neural responses, as described for adults (E. F. Chang et al., 2010; Correia et al., 2015).

2.3.2. An invariant code for sub-syllabic components

Second, we examined the invariance of the neural code by training new sets of manner and place estimators on a single context (e.g. stimuli spoken by the female voice) and testing them on the alternative untrained condition (e.g. male voice). We considered the speaker context in a first analysis and the vowel in a second analysis. Since several adult and infant studies have shown that information about phonemes and about speaker identity is encoded separately at an early processing stage (Formisano et al., 2008; Bristow et al., 2008), we expected full generalization across voice genders. As explained in the introduction, successful generalization across vowels would be indicative of sub-syllabic processing.

For manner, the timing of cross-context decoding was virtually identical to that seen in the overall analysis, and the accuracy only marginally reduced (Figure 2.3A, Table 2.1-2). Such generalization proves that the infant brain encodes manner features uniformly and irrespective of harmonic particularities, corroborating and extending previous behavioral evidence from older infants (Hillenbrand, James, 1983). Remarkably, clear generalization across voices and vowels was also obtained for place (Figure 2.3B). The time-course of classification, with two distinct decodable periods, and its accuracy were comparable to those achieved in the initial analysis (Figure 2.3B, Table 2.1-2). Since the acoustic cues for place vary substantially with the context (Liberman et al., 1967; Dorman et al., 1977), these cross-condition performances clearly reveal that the infant brain is able to extract an invariant code beyond acoustic differences, even in the challenging case of place contrasts.

Complementarily to these results, vowel estimators trained on single manner or place conditions fully generalized to the alternative contexts (Figure 2.3C and Table 2.1). Thus, the cross-decoding patterns observed so far demonstrate that syllables are not perceived holistically but are broken down into sub-components independently of the co-articulated vowel for consonants, and consonantal features for vowels.

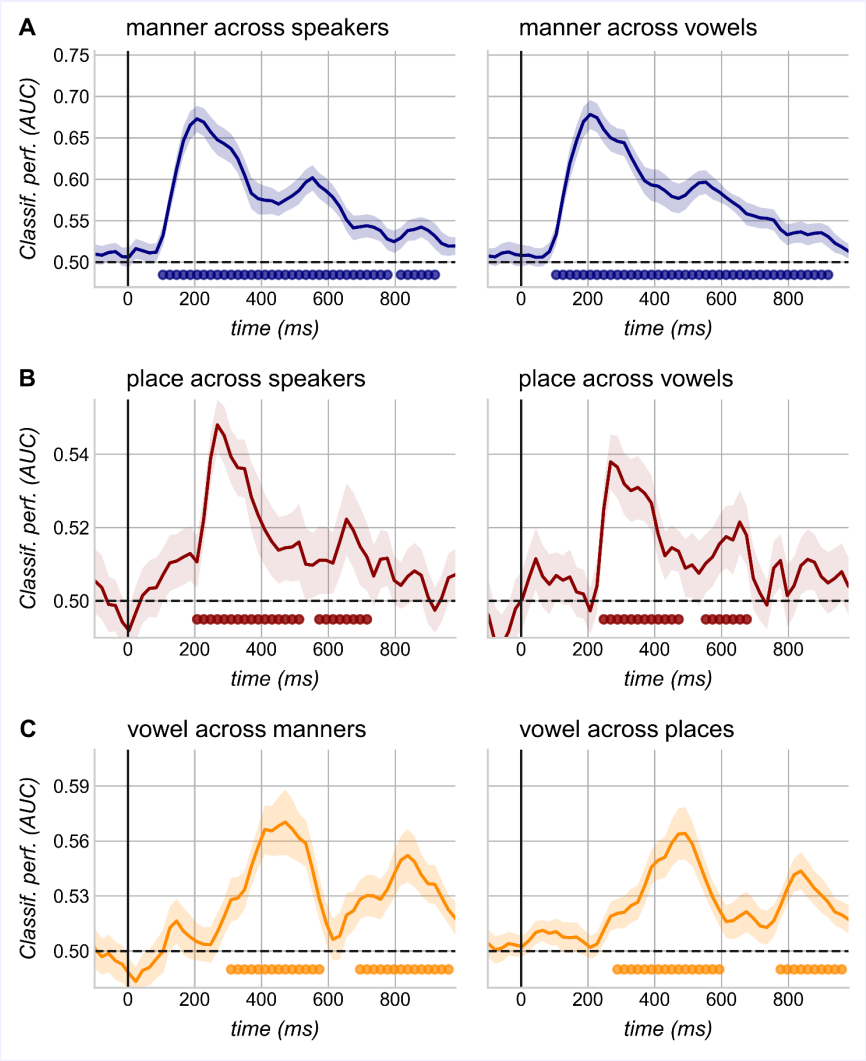

Figure 2.3 Cross-condition decoding.
(A) Left: generalization of manner estimators across voice conditions: classifiers trained on syllables produced by one speaker are tested on stimuli uttered by the other speaker. Right: generalization of manner estimators across vowel conditions: classifiers trained on consonants associated to one vowel are tested on syllables containing the alternative vowel. (B) Same as A, but for place estimators. (C) Left - vowel classification across manners: classifiers are trained on obstruents then tested on sonorants and vice versa. Right - vowel classification across places: vowel estimators are trained on one place condition (e.g. labials) and tested on the other two (e.g. alveolars and velars).
Shaded areas correspond to the standard error (SEM) across subjects; dotted black lines mark theoretical chance level. Filled circles indicate scores significantly above-chance (exact p-values are reported in Table 1). Performances from all possible training/test directions are averaged.

Classes based on:	generalization across:	*time window (ms)*	*p-clust*	peak performance		
				latency (ms)	*score*	*SD*
manner	speakers	100-920	0.0001	200	0.673	0.079
	vowels	100-920	0.0001	200	0.678	0.086
place	speakers	200-520	0.0001	260	0.548	0.035
		560-720	0.0014	640	0.522	0.047
	vowels	240-480	0.0001	260	0.538	0.034
		540-680	0.006	640	0.522	0.042
vowel	speakers	260-580	0.0001	460	0.561	0.078
		760-920	0.0002	800	0.554	0.052
	manners	300-580	0.0001	460	0.57	0.08
		680-960	0.0001	820	0.552	0.067
	places	280-600	0.0001	480	0.564	0.082
		760-960	0.0001	820	0.544	0.066

Table 2.1: Cross-condition decoding Summary of the decoding performances shown in Figure 2.3.

2.3.3. Syllables are first factorized into orthogonal codes corresponding to place and manner features, which are secondarily integrated

Holistic, unrelated codes for each of the six consonants might suffice for classifiers to sort trials into arbitrary subsets (e.g. /b/,/d/,/g/ vs /m/,/n/,/ɲ/), as shown in the previous sections. Crucially, if infants encode consonants by factorizing them into separate orthogonal dimensions, akin to the phonetic features postulated by linguists, then successful generalization should be obtained for decoders trained on one featural dimension, regardless of the variation in the other phonetic domains. That is to say, estimators would retrieve the same manner code across labials, velars and alveolars and the same place code in obstruents as in sonorants. To evaluate this possibility, we trained decoders in one featural context (e.g. manner classifiers were trained only on labials) and tested them on left-out data either *within* the same context (labials) or *across* untrained phonetic contexts (e.g. alveolars or velars). According to the decomposition/factorized hypothesis the two tests should yield similar performances.

This criterion revealed two distinct stages (Figure 2.4A): during an early time-window, both manner and place estimators achieved successful generalization, with a classification accuracy approaching that obtained within the trained condition. Initial processing was therefore based on orthogonal codes for the dimensions of manner and place. Beyond ~450ms however, classification performance was significantly lower across contexts as compared to within, suggesting a change in the format. Cross-condition decoding fell to chance level for place, while manner information was more resilient

but nevertheless altered by the variation in place context (Figure 2.4A). This finding suggests that a second phase of processing involved the grouping of multiple elementary dimensions into an integrated neural code, i.e. during this later time window, features were merged and no longer encoded as orthogonal, separately decodable dimensions.

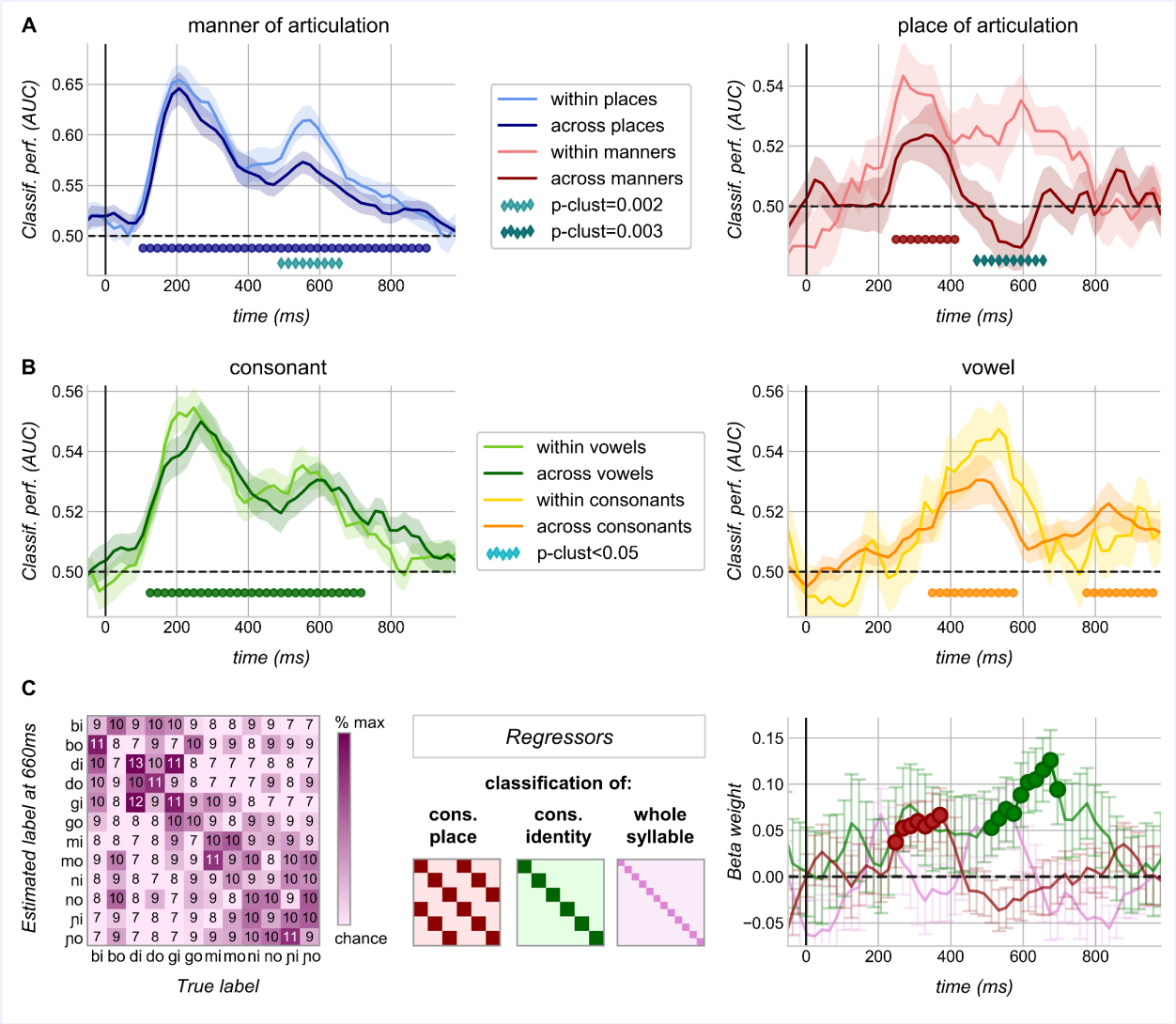

Figure 2.4 Orthogonal feature codes are merged into phoneme identities at a late stage of processing. (A) Time-resolved performance of estimators trained on a single phonetic feature (e.g. manner estimators trained on labials: /b/ vs. /m/). In light colors: classification within the trained condition (e.g. test on labials); in darker colors: performance at novel phonetic contexts (e.g. test on alveolars: /d/ vs. /n/, and velars: /g/ vs. /ŋ/). Scores from all possible training conditions or train/test directions are averaged. Shaded areas correspond to the SEM across subjects. Filled circles indicate significant generalization across contexts (100-900ms: p_{clust}=0.0001 for manner; 240-420ms: p_{clust}=0.001 for place). Diamonds indicate higher performance *within* as compared to *across* conditions (exact time window of significance for manner: 480-640ms; for place: 460-660ms). (B) Left: performance of estimators trained on discriminating all consonants (/b/ vs /d/ vs /g/ vs /m/ vs /n/ vs /ŋ/) coupled with one vowel (e.g. "-i") and tested within the same

(light green) and across the other vocalic context (e.g. "-o"; dark green). Right: performance of vowel classifiers trained on a single consonant (e.g. /b/) and tested within the same consonant (yellow) and across the remaining five (orange). Filled circles mark significant generalization across contexts (consonant classifiers: 80-900ms, p_{clust}=0.0001; vowel classifiers: 340-560ms, p_{clust}=0.0001 and 760ms onwards, p_{clust}=0.0002).
(C) Left: example of a neural confusion matrix at time t (660ms) obtained with a 12-class (syllables) decoding problem (average across subjects). Numbers within each cell indicate the percentage of times a given syllable from the x-axis was classified with the label reported on the y-axis. Off-diagonal values diverging from 0 signal misidentification (chance=8.3%). Middle: theoretical confusion matrices depicting a perfect separation between (i.e. the ideal classification of) consonantal places, consonant identities and broad syllable identities (classes are ordered as in the left matrix). Darker colors correspond to the values 25%, 50% and 100% respectively, light colors correspond to 0%. These matrices were entered as predictors of interest in a multiple regression analysis to explain neural syllable confusion at each time point. Right: the obtained beta-weights averaged across subjects and marked by filled circles when significantly above zero (cluster-based permutation t-test). Vertical lines correspond to SEM. To enhance clarity, the remaining predictors (i.e. manner and vowel discrimination) and the relative beta-weights are illustrated in section 2.5, Figure 2.9B.

2.3.4. Consonant and vowels remain separated

Were the consonant and the vowel ever merged in a syllabic unit? The results obtained so far contain a few interesting hints in this regard. As shown in Figures 2.2 and 3.2, vowel decodability follows a double-peak pattern very similar to that observed for consonantal dimensions, but peak scores are achieved markedly later and at times when consonantal place is hardly discriminable. Together with the invariance of vowel codes across consonantal features (Figure 2.3C), these observations suggest that infants encoded the two phonemes composing the syllable in a separate and well-ordered fashion.
In a final step, we queried a possible interconnection between consonant and vowel processing. Using a logic similar to the one described above, we compared the performance of consonant and vowel estimators *within* and *across* vowel and consonant conditions. The presence of an integrated syllabic code would generate a drop in performance across context. As displayed in Figure 2.4B, such drop never occurred, suggesting that consonants and vowels were kept separated, at least until 1 second after syllable onset.

All the decoding results described above were further validated by the sanity check analyses illustrated in Figure 2.8 (section 2.5), where we used randomized training sets and arbitrary cross-condition tests. By showing the absence of haphazard decodability, the latter confirmed [a] the appropriateness of the stimuli set employed; [b] the reliability and interpretability of the multivariate techniques applied; [c] the non-arbitrariness of phonemes and phonetic features as relevant linguistic dimensions.

2.3.5. Neural confusion matrices

To gain additional evidence on the nature of the encoding across time, we trained algorithms on whole syllable identities (i.e. 12 labels: “bi” vs “bo” vs ”di” vs ”do” vs ”gi” vs “go” etc.) and explored their error patterns at test. With this decoding scheme, class separation might be based on either one or a mixture of the stimuli dimensions explored so far. It follows that, in this analysis, class-wise accuracy (Figure 2.9A-top) will be poorly informative per-se. Between-class confusion, on the other hand, can provide an exhaustive picture of the encoding modality at each time point. For instance, whereas the retrieval of neural codes for whole syllables would produce a purely diagonal confusion matrix, phoneme-identity neural codes would trigger conspicuous mislabeling among pairs of stimuli sharing the same consonant or vowel. Using multiple linear regression, we tested whether and when pairwise neural syllable confusion (Figure 2.4C-left and 2.9A-bottom) was explained by the isolation of either featural, consonant-identity and/or whole-syllable codes (Figure 2.4C-middle) once vowel distinctions were entered as a variable of non-interest (since our paradigm did not enable to disentangle vocalic features from vowel identity). We found that consonantal place of articulation drove neural confusability early in the trial (240-380ms: p_{clust}=0.017). Crucially, consonant-identity predicted the patterns of neural separability *only* later, between 500 and 700ms (Figure 2.4C-right; p_{clust} = 0.006). Lastly, the syllable regressor never reached significance (Figure 2.4C). Complementing the decoding outcomes in Figure 2.4A-B, these results show that following the encoding of orthogonal features, place and manner codes were integrated into comprehensive consonant bundles, while consonants and vowels remained separated.

2.4. DISCUSSION

The classification patterns observed in this study reveal two speech encoding formats in the infant brain. During a first stage of processing, each consonant was encoded by its coordinates along the manner and place dimensions, as evidenced from the fact that decoders trained on one dimension could generalize to different levels of the other dimension. In a second stage, the two features were combined into idiosyncratic bundles, still allowing phoneme classification but hindering full generalization of featural decoding across different consonants. This functional progression is consistent with the dynamic nature of the neural codes as revealed by the matrices in Figure 2.2D and the corresponding informative activity patterns in Figures 2.6-7. Although our experiment was mainly focused on consonants, similar processing stages for vowels are likely. Finally, we found no evidence for an encoding of the syllable in its entirety.

According to several mainstream accounts, authentic adult-like phonetic perception requires the acquisition of refined motor skills that would enable a proficient mapping between articulatory movements and acoustic outcomes (Kuhl et al., 2008; Schwartz et al., 2012; Vilain et al., 2019; Westermann & Reck Miranda, 2004). Through vocal plays, aimed at imitating ambient language, infants would gradually familiarize with the sensory consequences of their own utterances. Once

they begin to master production, the acquired availability of internal motor models would enable them to process speech sounds in phonetic terms (Kuhl et al., 2008, 2014; Vilain et al., 2019). In this scenario, canonical babbling, which signals the beginning of a fairly controlled articulation around 6-8 months of age (van der Stelt & Koopmans-van Beinum, 1986), represents an important milestone, while infants in the pre-babbling phase are thought to rely on refined but domain-general auditory mechanisms (Kuhl, 2004). It follows that, according to these widely accepted views, the primitive units for speech processing consist of spectrotemporally detailed but phonetically undefined acoustic chunks roughly corresponding to syllables.

The decoding performances shown here suggest a different developmental scenario. First, the observed separation between consonants and vowels demonstrates that, even for pre-babbling infants, syllables are not holistic units. Without diminishing the importance of syllabic-level analysis (e.g. 42), our finding of neural codes for consonant identity complements adult data (Zhang et al., 2016) in corroborating the reality of the phoneme as a relevant entity for the cortical encoding of speech (Kazanina et al., 2018).

Second, our generalization approach, involving the comparison of decoding performances within- and across-phonetic domains, disclosed the existence of a preliminary phase where consonants are decomposed along distinct and orthogonal axes for the manner and place of articulation. Although we tested only two consonantal features, the characteristics of our experimental design allow strong insights upon the nature of such first encoding stage. To start with, we carefully selected the stimuli to avoid any trivial difference, for instance in consonant duration (see section 2.2.2, *Stimuli construction*). Importantly, we opted for the dimensions of place and manner because the consistency of their acoustic correlates across contexts is largely different. Further, the experimental stimuli were appositely chosen to push the variability of place cues at the maximum (e.g. /i/ vs /o/, situated at opposite corners of the vowel diagram, accentuated the spectro-acoustical inconsistency of place cues due to co-articulatory phenomena (Liberman et al., 1967). Such *a-prioris* were confirmed by our inspection of the auditory spectrograms (Figure 2.5) where the acoustic similarity between tokens was explained by manner, vowel and voice commonalities but not place. Yet, on EEG recordings, cross-classification performances for both features remained qualitatively similar and disclosed invariant neural codes that outreach context-dependent spectrotemporal details. These observations suggest that, within a first stage of processing, the infant brain is capable of reducing the intrinsic sensory richness of the speech input by factorizing it. In this fashion, a complex signal, varying along many axes, is compressed by projection onto a few, linguistically relevant, dimensions.

Overall, the current study shows that the neural foundations of speech perception are strikingly similar in infants and adults (Mesgarani et al., 2014; E. F. Chang et al., 2010; Correia et al., 2015; Zhang et al., 2016; Khalighinejad et al., 2017), and compatible with the decomposition into distinctive features postulated by linguists (Halle, 2013). Other than providing evidence for phonetic encoding in pre-babblers, our results clarify some ambiguities from previous adult studies and extend our knowledge of human speech perception. In adults who passively listened to sentences, the EEG

revealed a temporal progression of phoneme-related potentials characterized by distinct topographies over a period ranging 50-400ms relative to phoneme onset (Khalighinejad et al., 2017). However, the experimental design did not allow to explore the functional significance of such evolving activity patterns. Cortical recordings in adults have uncovered that distinct electrodes encode different dimensions of the speech signal (Mesgarani et al., 2014), but they primarily observed the neural correlates of manner and voice-onset-time. Since the latter have clear acoustical signatures in the stimulus spectrum, such evidence might not suffice to conclude in favor of a genuinely featural code for speech. Meanwhile with fMRI, a multivariate decoding procedure equatable with that proposed here allowed to uncover feature-specific responses in various areas of the adult temporal lobe (Correia et al., 2015). Our findings are fully congruent with all these observations carried on subjects who master their native language, thus supporting a continuity in speech encoding from the learner to the expert. Furthermore, our results unify these previous insights into a coherent picture: we propose that the extraction of minimal orthogonal features (Correia et al., 2015), constitutes the first step of a perceptual *process* (Khalighinejad et al., 2017) leading to phoneme identity computation. Such a process creates a structured and highly generalizable space that is robust to surface variability across speakers and co-articulatory contexts.

A factorized representational mechanism was previously discovered in the monkey face patch system (L. Chang & Tsao, 2017). As outlined for the visual domain, such a decomposition strategy applied to speech is more parsimonious, efficient and flexible than exemplar coding (e.g. R. Port, 2007; R. F. Port, 2010). Given these characteristics, a factorized encoding system seems ideally suited to bootstrap learning: it enables infants to discover linguistic regularities based on the combinatorial possibilities of a reduced set of elements rather than a large diversity of syllables and spectro-temporal patterns.

In particular, a code based on invariant phonetic features might play a crucial role in lexicon acquisition. A first support for this claim comes from evidence demonstrating its effectiveness in real-world scenarios: when minimal phonetic distinctions are embedded in acoustically prominent but irrelevant variations, infants become especially prone to catch phonetic regularities in order to learn words (Rost & McMurray, 2009). In this context, the vectorized system we propose discards the irrelevant variability to organize the input according to phonetic criteria; such perceptual re-organization turns up those subtle phonetic differences that define word's meaning. Importantly, in order to discover words, infants must cope not only with acoustical but also with phonological variation due to the segmental context: for example, in order to apprehend that "wet shoes" and "we[p] pants" share the same word "wet", English infants should apply a rule stating that an alveolar stop consonant borrows the place of articulation from the subsequent stop (Darcy et al., 2009). Phonotactic rules of this sort pertain to phonetic features rather than holistic phonemes. Several behavioral studies reported that infants are sensitive to phonotactic cues already by the age of 9 months: they prefer to listen to sequences that are phonotactically legal in their native language (Friederici & Wessels, 1993; Jusczyk et al., 1993) and use their phonotactic knowledge to find word

boundaries in continuous speech (Mattys & Jusczyk, 2001). At this age, coherently with our argument, phonotactic rules are easily learned if expressed at the level of phonetic features while they are not detected when they concern the identity of the phonemes (Saffran & Thiessen, 2003). Lastly, a featural encoding of speech is consistent with the documented ability of young infants to use phonetic details in word-referent mapping (Swingley & Aslin, 2002; Fennell & Waxman, 2010).
Also the neural separation between consonants and vowels, which characterizes the second stage of processing, seems particularly valuable for learning. Consonants and vowels have been proposed to hold diverging roles in language: while the former carry lexical distinctions, the latter are especially apt to mark structural organization (Nespor et al., 2003). Their encoding as orthogonal/separate entities enables the maintenance of two parallel pathways of processing, optimizing in this way the accessibility of lexicon on one side and syntax on the other. Coherently with our findings, and just as adults (Toro et al., 2008), infants are known to exploit the "division of labor" between consonants and vowels already by the age of 12 months (Hochmann et al., 2011). The inclusion of different syllabic structures in future experimental paradigms will bring further insights, e.g. investigations with CCV/CVV tokens will enable to elucidate whether orthogonal encoding concerns single phonemes or rather consonantal/vocalic functional clusters.

Phonetic features and phonemes might then correspond to essential and quickly available building blocks for human language acquisition. Still, the developmental origin of these codes, and in particular their dependence on motor representations, require further study to be understood. At ~12 weeks, the age of our subjects, vocal production is very limited (Kuhl, 2004). Strikingly, even preterm neonates can detect a place of articulation change ("ba" vs "ga") at 6 months of gestation, when articulatory movements are extremely poor. Before term, such discriminative ability is carried by a network of temporal and frontal brain areas similar to that recruited at later ages (Mahmoudzadeh et al., 2013, 2016). These observations suggest that the encoding system isolated here develops prior to, and independently of, motor skills. Nevertheless, orofacial stereotypies such as tongue protrusion/retraction occur already in the womb and protophones, the earliest precursors of oral language, start to be produced, in an exploratory fashion, immediately after birth (Oller et al., 2019). These primitive behaviors could provide a primordial knowledge of the shape and configurability of the upper vocal tract (Choi et al., 2017) and, combined with sound exposure, they might foster an integrative/multi-modal representational space for speech before the onset of canonical babbling. Coherently with this conjecture, a recent study in 3-month-olds showed that altering the movements of the tip of the tongue modulates the perception of a labial-alveolar contrast, thereby revealing the presence of a refined auditory-motor mapping (Choi et al., 2021). Although multi-modal speech processing appears from an early age (Bristow et al., 2008), the perceptual stage at which different modalities are integrated, as well as their relative weights, remain to be determined.

As a final remark, we would like to warn the reader about two interpretative issues our methodology entails. Strictly speaking, our multivariate decoding approach revealed a statistical dependence between a psycholinguistically-defined representational space composed of phonetic vectors and the spatiotemporal activity patterns captured by the EEG sensors (Hebart & Baker, 2018; Kriegeskorte & Bandettini, 2007). When conceiving the brain as an information-processing system based on population coding (Panzeri et al., 2015), pattern-information analyses are likely to have considerable functional significance, especially in comparison to more classical activation-based approaches (Hebart & Baker, 2018; Kriegeskorte & Bandettini, 2007). Furthermore, our choice of linear (as opposed to non-linear) classifiers ensures the biological plausibility of our conclusions (Kriegeskorte, 2011). Nonetheless, demonstrating that neural activity patterns incorporate phonetic information does not necessarily imply that the infant brain actually uses such information for its operations. The literature provides two hints in this direction. First, a behavioral investigation relying on the head-turn preference procedure reported that 4-month-olds could successfully learn a phonotactic rule shaping vowel-consonant parings on the basis of featural classes (i.e. "nasal vowels are always followed by fricatives, and oral vowels by stop consonants")(Seidl et al., 2009). Moreover, a recent ERP study found that when exposed to syllables varying in their vocalic constituents, 3-month-old infants could learn to pair consonants and visual shapes and generalize this pairing to a new vocalic context, demonstrating that sub-syllabic representations are already operational at this age (Mersad et al., 2021). We point this line of study as a meaningful direction for future research.
A second interpretative issue might arise from linear models being, by nature, strongly dependent on the experimenter's a priori insights: by fitting only the phonetic variables included in our hypothesis, we might have missed the influence of unexpected variables possibly accounting for the successful classification of the former. In light of such caveat, the emergence of phonetic codes in a (relatively) unsupervised decoding analysis is particularly noteworthy (Figure 2.4C and 2.9). Namely, in absence of any predefined stimulus grouping, the representational structure revealed by the confusion patterns of syllable classifiers matched the predictions of the phonetic representational space hypothesized.

To conclude, pending more definitive experimental evidence, we point out the possibility that an abstract, combinatorial code for speech might be available very early on and endow infants with the ability to discriminate phonemes from most languages (Jusczyk, 2000). We further highlight that an encoding system based on a finite set of minimal and orthogonal elements is ideally suited to bootstrap the acquisition of phonotactic, lexical and syntactic rules. The method presented here provides the foundation for future experiments that, spanning a range of languages and ages, will need to investigate how the observed codes develop and adapt to the inventory of native phonemes.

2.5. SUPPLEMENTARY MATERIALS

2.5.1. Auditory spectrogram estimation and Representation Similarity Analysis

This preliminary investigation was aimed at delineating the auditory representational geometry elicited by our stimuli set (Kriegeskorte, 2008; Kriegeskorte & Kievit, 2013).

The time-frequency auditory representation of the speech sounds was estimated according to a model of the peripheral auditory system (Chi et al., 2005) as implemented in the NSL Matlab Toolbox (http://nsl.isr.umd.edu/downloads.html). This model comprises: a first step in which sound frequencies are spatially separated along the basilar membrane; a second stage that simulates the transduction of basilar membrane displacements into auditory nerve spikes; and a third phase of processing within the cochlear nucleus. The output of the model is an auditory spectrum of the signal as it enters the inferior colliculi. The three stages and their mathematical implementations are described in (Yang et al., 1992) and (Wang & Shamma, 1994). Auditory spectra were computed based on consecutive windows of 10ms for each stimulus, obtaining a total of 120 bidimensional (time *x* frequency) auditory representations. We then estimated pair-wise auditory dissimilarity following two different approaches.

First, we calculated time-resolved auditory (dis)similarity. For this purpose, spectrograms were aligned upon the consonant offset times determined with the gating procedure described in the Materials and Methods (section 2.2.2). Consonant offset was preferred over syllable onset because acoustic cues for the place of articulation are generally proposed to reside within the formant transitions (i.e. at the time of the switch between consonant and vowel portions) (Liberman et al., 1954). Since consonant duration varied across speech tokens, alignment based on syllable onset would have led to a jittering of such transition times across spectrograms and this jittering could have misleadingly attenuated relevant cues. The 5 auditory spectrograms corresponding to each sub-condition (e.g. the 5 utterances of "go-female") were then averaged together (Figure 2.5B). For each (10ms long) spectral frame, we z-scored amplitude values across frequencies and calculated the Euclidean distance between each pair of sub-conditions. Standardization was applied in order to maximize our power of detecting phonetic distinctions despite variation in fundamental frequencies (i.e. despite male and female voices being characterized by very distinct pitches). The choice of the Euclidean metric is justified by its potentiality to mimic infant discriminative behavior with higher fidelity relative to other distance measures (Sundara et al., 2018). The outcome of this first approach is a series of 35 auditory distance matrices (Figure 2.5B), describing all together how pairwise auditory (dis)similarity unfolds over time.

It has been proposed that the acoustic correlates of the place of articulation, a feature of major interest in the current study, have an integrative and dynamic nature (Nossair & Zahorian, 1991). The employment of brief time slices could have then potentially precluded us from capturing meaningful cues derivable from the spectral shape as a whole. To account for this eventuality, our

second approach relied on the Dynamic Time Warping (DTW) algorithm (Sakoe & Chiba, 1978; Park & Glass, 2008) as implemented in the Python module *dtaidistance* (Meert & Van Craenendonck, 2018). This technique enabled us to find the best alignment between each pair of spectrograms by stretching and compressing them locally, along the time axis. Following z-scoring, we estimated the DTW distance between each pair of utterances and obtained a comprehensive auditory dissimilarity matrix by averaging the distance values corresponding to each pair of sub-conditions.

To investigate the relationship between the auditory space and the phonetic/harmonic dimensions of our speech stimuli we tested the correlation of the auditory distance matrices with four theoretical matrices (Figure 2.5C). The latter consisted of categorical models in which two syllables are identical (dissimilarity = 0) if they share the same manner/place/vowel/voice, and different (dissimilarity=1) in case they do not. Concerning place of articulation distinctions, some investigations in phonetics seem to suggest that labials/velars and alveolars could be acoustically closer to each other relative to labial and velars (Cho & Ladefoged, 1999; Lisker & Abramson, 1964). Furthermore it has been proposed that the alveolar feature may be "underspecified" (i.e. coronal may correspond to the default place and therefore be somehow inactive/less contrastive) as compared to the labial or velar features (Cummings et al., 2017; Stemberger & Stoel-gammon, 1991; Tsuji et al., 2015). To account for these possibilities, we built an additional model where the distance between labials and alveolars and that between alveolars and velars was quantified as "0.5". Results obtained with the two place models were completely overlapping.

The match between auditory and theoretical dissimilarity matrices was quantified with a Mantel test for two-dimensional correlations (Mantel, 1967) employing Spearman's rho as test statistic and performing 10000 permutations for each test. The Mantel procedure, unlike the classical correlation methods, enabled to account for the fact that distances here were not independent, i.e. every dissimilarity depended on two spectral patterns/qualitative values, each of which also codetermined the similarities of all its other pairings in the matrix. False discovery rate (FDR) correction was applied in case of multiple comparisons across spectral frames.

Time-resolved outcomes are show in Figure 2.5D. Along an average sound duration of 400ms, the auditory pairwise dissimilarity of the stimuli is best described by manner of articulation distinctions up to 140ms (i.e. during the consonantal portion) and later by the vowel. Acoustic similarities are additionally shaped by voice gender throughout the entire syllable, while they do not have any straightforward relationship with the place of articulation (Figure 2.5D). The comprehensive auditory dissimilarity matrix is significantly correlated with manner (Mantel r_s =0.228, p=0.0002); vowel (Mantel r_s =0.297, p=0.0001) and speaker distinctions (Mantel r_s =0.24, p=0.0001) but not place of articulation (Mantel r_s =-0.029, p=0.75).

These results are coherent with the fact that, despite 70 years of research, investigators could not find an acoustic description of the place of articulation that is valid for all contexts. Intriguingly enough, although able to form place-based categories, animals have been shown to process place contrasts in a context-dependent way (Sinnott & Gilmore, 2004). Given all these elements, the ability

to detect stable place contrasts across different production circumstances might be considered as the ultimate challenge to address in order to understand human speech perception.

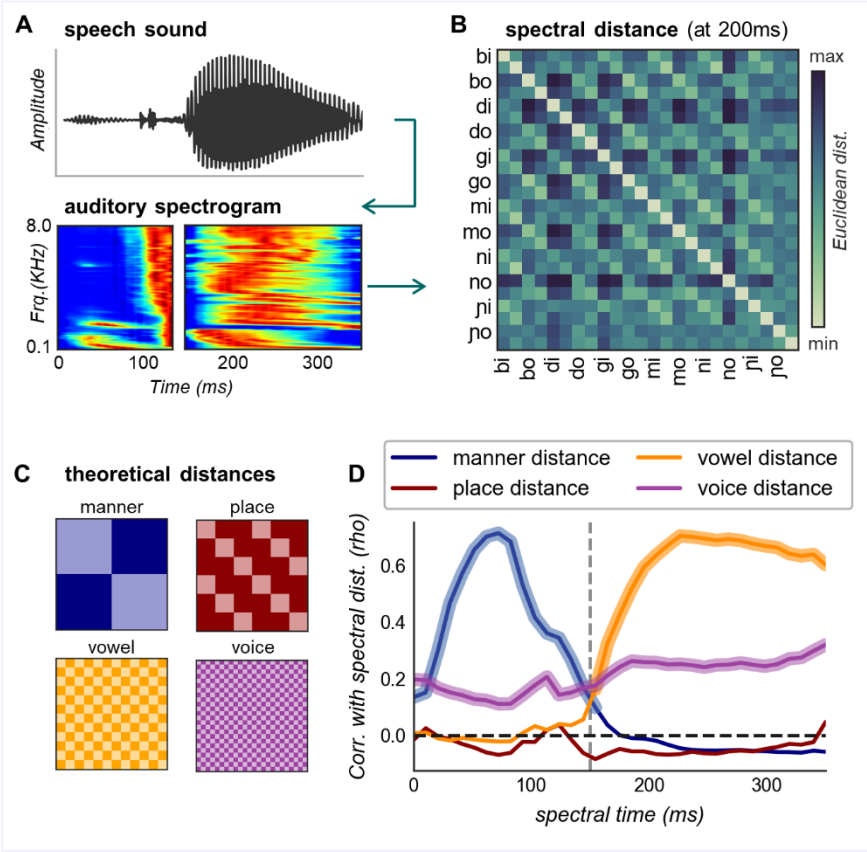

Figure 2.5 Representational content of the speech stimuli (Figure 2.1) as they reach the central auditory pathways.
(A) Auditory spectrograms were extracted from the speech sounds with a model of cochlear frequency analysis, then averaged by syllable type (top: one instance of "go" pronounced by the female voice; bottom: average spectrogram of all 5 utterances belonging to the sub-condition "go-female"). The blue-red scale reflects minimal-maximal energy, separately normalized in the consonant and vowel portions for mere illustrative purposes. (B) Example of dissimilarity matrix reporting the Euclidean distance between each pair of auditory spectrograms at spectral time=200ms. Each label (e.g. "bi") indexes two sub-conditions: female and male. (C) Categorical dissimilarity models (conditions are ordered as in the matrix above): light colors indicate correspondence (distance=0) while darker colors signify lack of correspondence (distance=1). (D) Correlation between spectral and theoretical distance matrices as syllable unfolds (the dotted vertical line marks the switch between consonant and vowel). Thicker lines indicate significant time points ($p<0.05$) after FDR correction. Full methodology description, rationale and complementary results are reported in the supplementary text above.

As a note, the reader may wonder the reason why we could not apply the same decoding strategies used on neural data in order to characterize the auditory space. Generally speaking, the lower the number of samples and the higher the ratio of features to sample size, the more a machine learning model will fit the noise in the data instead of a meaningful pattern (Jain & Chandrasekaran, 1982;

Kanal & Chandrasekaran, 1971). In the case of our auditory spectrograms, algorithms would need to be trained/tested on a maximum of 120 samples with 4480 features each (as a benchmark: samples for each neural estimator in the main analyses were approximately 1600 and contained 1260 features each). Evidently, such disproportionate dataset is ill-suited for the same kind of estimators used on the ERPs: instability and overfitting would completely undermine the reliability (and therefore interpretability) of the outcome. On the other hand, the same RSA approach used to characterize the auditory space would have been a largely suboptimal strategy if applied on neural data, for the following reason. The signal-to-noise ratio in infant EEG is considerably worse than that of spectrograms (which, as described above, are estimated through a software starting from .wav files recorded in optimal environmental conditions). When provided with an adequate number of samples, machine-learning methods can overcome noise-related limitations by combining information from different EEG sensors, leading to a gain in sensitivity that could not be achieved by averaging individual trials together (Hebart & Baker, 2018). Nevertheless, we did perform a similarity-based analysis on neural data that parallels, at the conceptual level, the present investigation on auditory spectrograms. Namely, we quantified the neural (dis)similarity between syllabic conditions as the degree of confusion yielded by classifiers trained using generic labels (Figure 4C and 2.9). Note that, in this case, decoding was conducted in a relatively unsupervised fashion, resulting in neural confusability patterns that are qualitatively comparable with the acoustic dissimilarities. Of particular interest is that, in sharp contrast to the results obtained for the auditory space, place of articulation was a significant factor driving the (dis)similarity of the neural responses.

2.5.2. Weights projection

The weights assigned by classifiers to EEG sensors reflect the degree to which the information captured by a given sensor is used to maximize class separation. However, weights per se are very difficult to interpret. For example, higher weights do not necessarily correspond to high levels of class-specific information as they could be assigned to sensors that are employed to delineate and suppress noise (for a full explanation see: Haufe et al., 2014). To overcome this issue it is possible to project weights back onto an interpretable activation space by multiplying them with the covariance in the data (cov(X), where X is the N × M matrix of EEG data with N trials and M channels). In the resulting vector (that has length M channels) large amplitudes indicate high degrees of class-specific brain activity (Grootswagers et al., 2016; Haufe et al., 2014). Since our goal was to reconstruct informative activity peculiar to each phonetic feature domain, we retrieved the coefficients of classifiers trained within each place condition to obtain "pure" manner-distinctive patterns and trained within each manner condition to obtain "pure" place-distinctive patterns. By doing so, we ensured that no information about place was available to manner estimators and no information about manner was available to place estimators. After multiplying coefficients and EEG covariance, the resulting activity estimations were averaged across places (to obtain informative activity for manner) or manners (to obtain informative activity for place).

To identify sensors that were crucial specifically for manner or crucial specifically for place classification, we computed the 10th and 90th percentiles of the informative activity values observed throughout the trial. At each time point, channels whose informative activity amplitude fell below the 10th or above the 90th percentiles in one phonetic domain but not the other were interpreted as particularly important to manner but not place classifiers or vice versa (Figure 2.6).

Further, we compared labial- , alveolar- and velar-specific patterns of informative activity with 1-way repeated measures ANOVA (Figure 2.7B). As done for the main analyses, we addressed the multiple comparisons problem with a permutation procedure based on spatio-temporal clusters. Neighboring elements that passed a threshold corresponding to a p-value of 0.01 were grouped together and their significance was computed by comparing cluster-level statistics to a null distribution of f-value sums created by drawing 10000 random permutations of the observed data. A cluster was considered as significant when its p-value was below 0.05. Since informative activity patterns are meaningful only in case of successful decoding (Haufe et al., 2014), differences were evaluated only during the two time windows when place classification was reliably above chance.

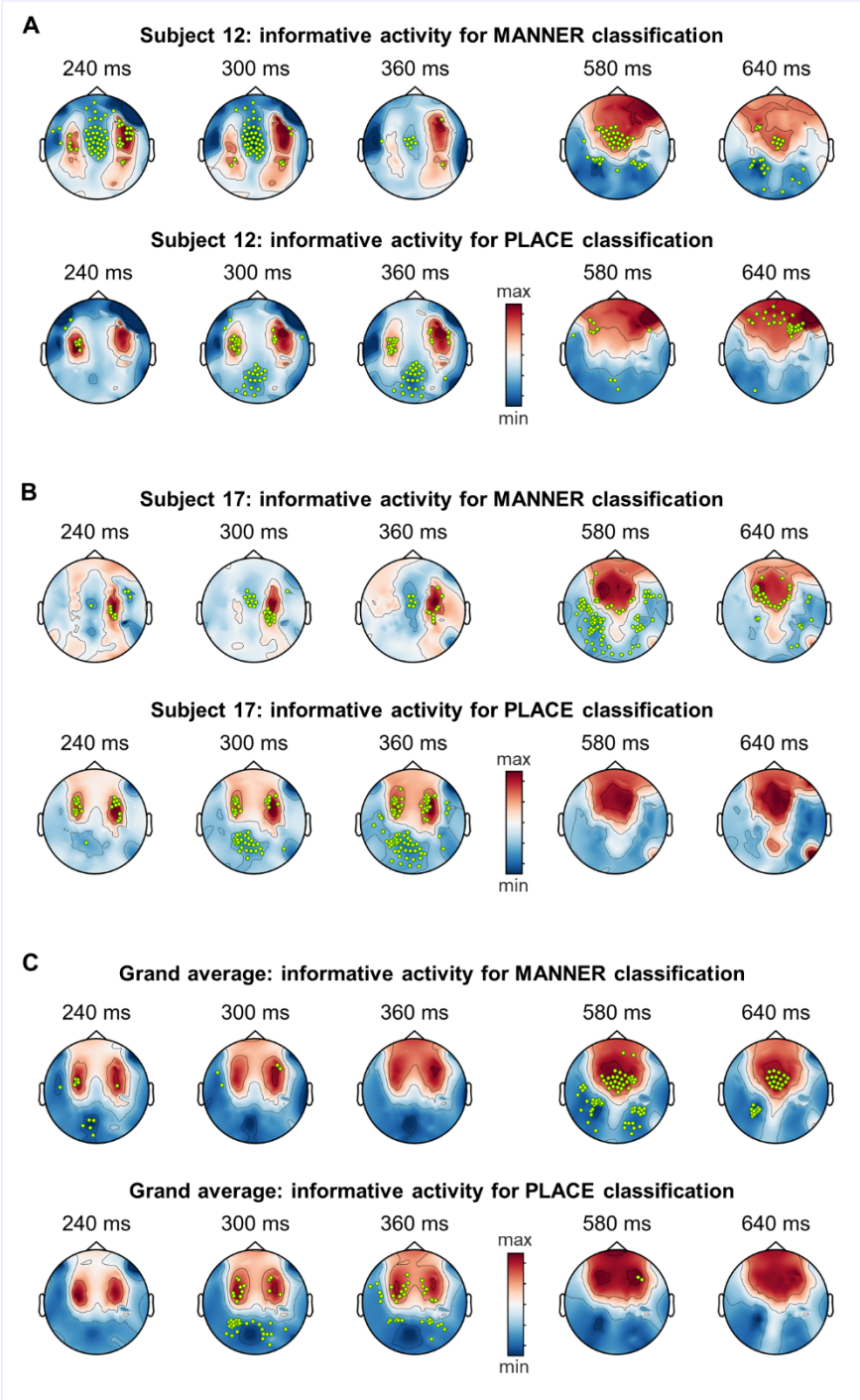

Figure 2.6 (complement of Figure 2.2) Discriminative loci change as a function of time and phonetic feature dimension

Classifiers weights are projected onto the EEG sensor activation space. Darker colors correspond to brain activity that was useful for classification. Marked in yellow are channels carrying crucial information to distinguish manner but not place (top rows) or to discriminate place but not manner (bottom rows). Time points are chosen to provide an overview of the two time-windows with reliable classification. Panels (A) and (B) show the informative activity patterns reconstructed for two representative subjects. In (C)

informative activity patterns are averaged across infants with the purpose of providing a visualization of the general trend. Note however that the interpretability of this grand average is limited since decoding analyses were carried within subject and discriminative loci are very much idiosyncratic. Overall, these topographies show that, as time passes, sensors conveying valuable information are located more medially over frontal areas. Moreover, informative locations for manner and place of articulation do not always overlap.

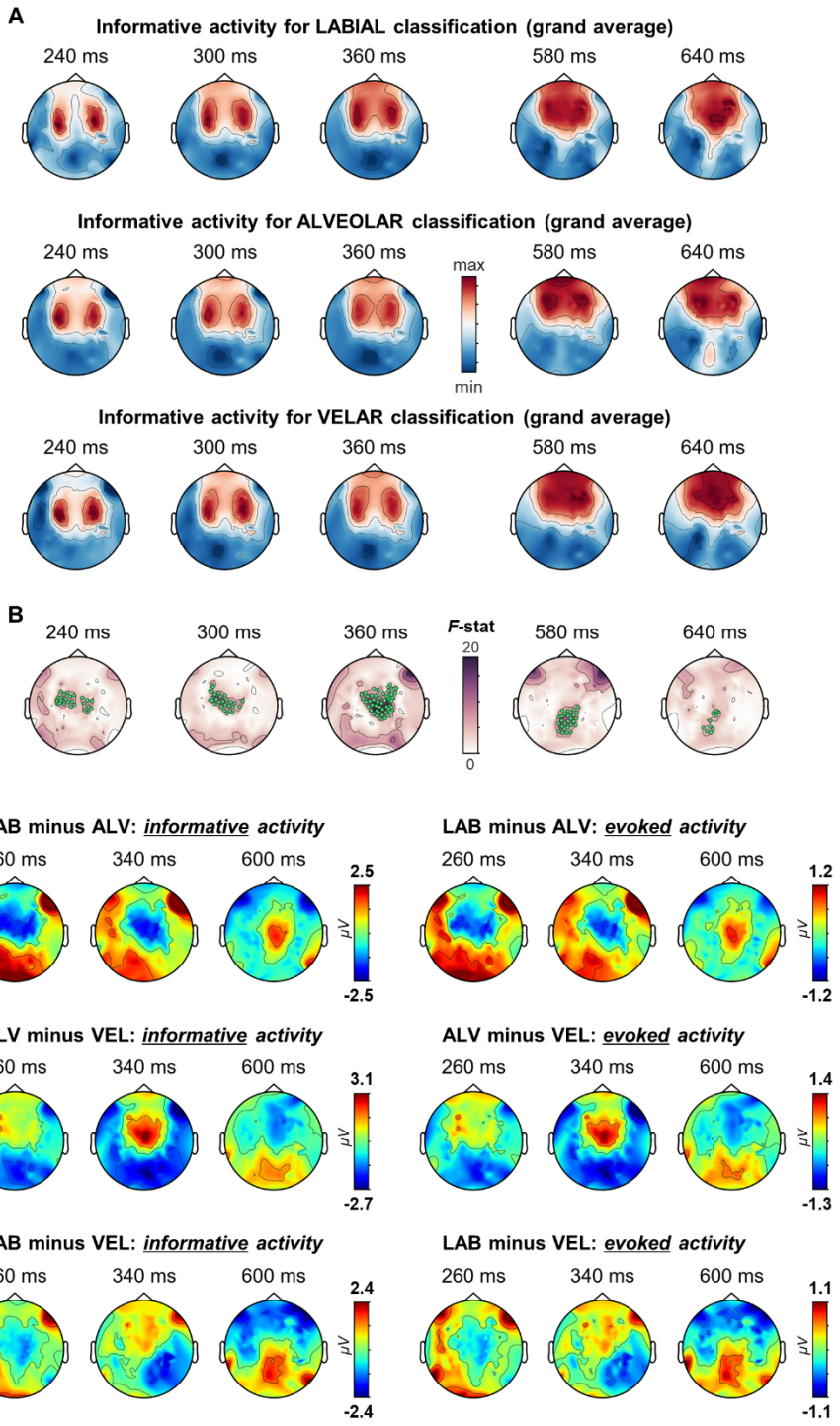

Figure 2.7 (related to Figure 2.2) Overview of place contrasts: informative and evoked activity patterns.

(A) In place decoding, three distinct models were fitted to separate each place of articulation from the other two (one-vs-rest approach). Their weights were projected back onto the activation space to reconstruct patterns of activity useful in characterizing either labials, alveolars or velars against the other places. Darker colors correspond to loci providing high degrees of class (i.e. place)-specific information. Patterns are averaged across subjects to provide an impression of the general trend; note however that weight idiosyncrasy undermines the interpretability of the grand average. (B) Results of one-way repeated measures ANOVA comparing discriminative activity for labials vs. alveolars vs. velars; channels containing significant differences are in green: early time-window: p_{clust} =0.0005, late time-window: p_{clust} =0.0196. (C) Reported on the left are differential informative activity patterns, on the right the same differences were computed on the evoked related potentials (ERPs). Given that amplitude ranges of informative and evoked brain activity were extremely similar (spanning from -8 to 7 μV in both cases), this figure displays two remarkable features: differential topographies are qualitatively overlapping while amplitude scales (colorbars) change substantially from the left to the right side of the panel.

phonetic feature	time window (ms)	decoding analysis	comparison to overall classification		
			mean score	*t(24)*	*p*
manner	200 - 400	overall	0.685±0.065		
		across genders	0.643±0.057	2.278	0.032
		across vowels	0.649±0.0523	2.176	0.040
place	260-360	overall	0.538±0.039		
		across genders	0.548±0.031	-1.085	0.289
		across vowels	0.536±0.033	0.194	0.848
	580-680	overall	0.526±0.039		
		across genders	0.513±0.041	1.353	0.189
		across vowels	0.519±0.032	0.651	0.521

Table 2.2. Formal comparison between main and cross-condition decoding of phonetic features Performance of estimators trained on exclusive conditions ("across"; Figure 3A-B) is compared to that of estimators trained on all conditions at once ("overall"; Figure 2A-B). AUC scores were averaged over 200ms (the first time point to be considered was set upon peak performance) and, once ascertained the normality of each distribution, contrasted with two-sided t-tests.

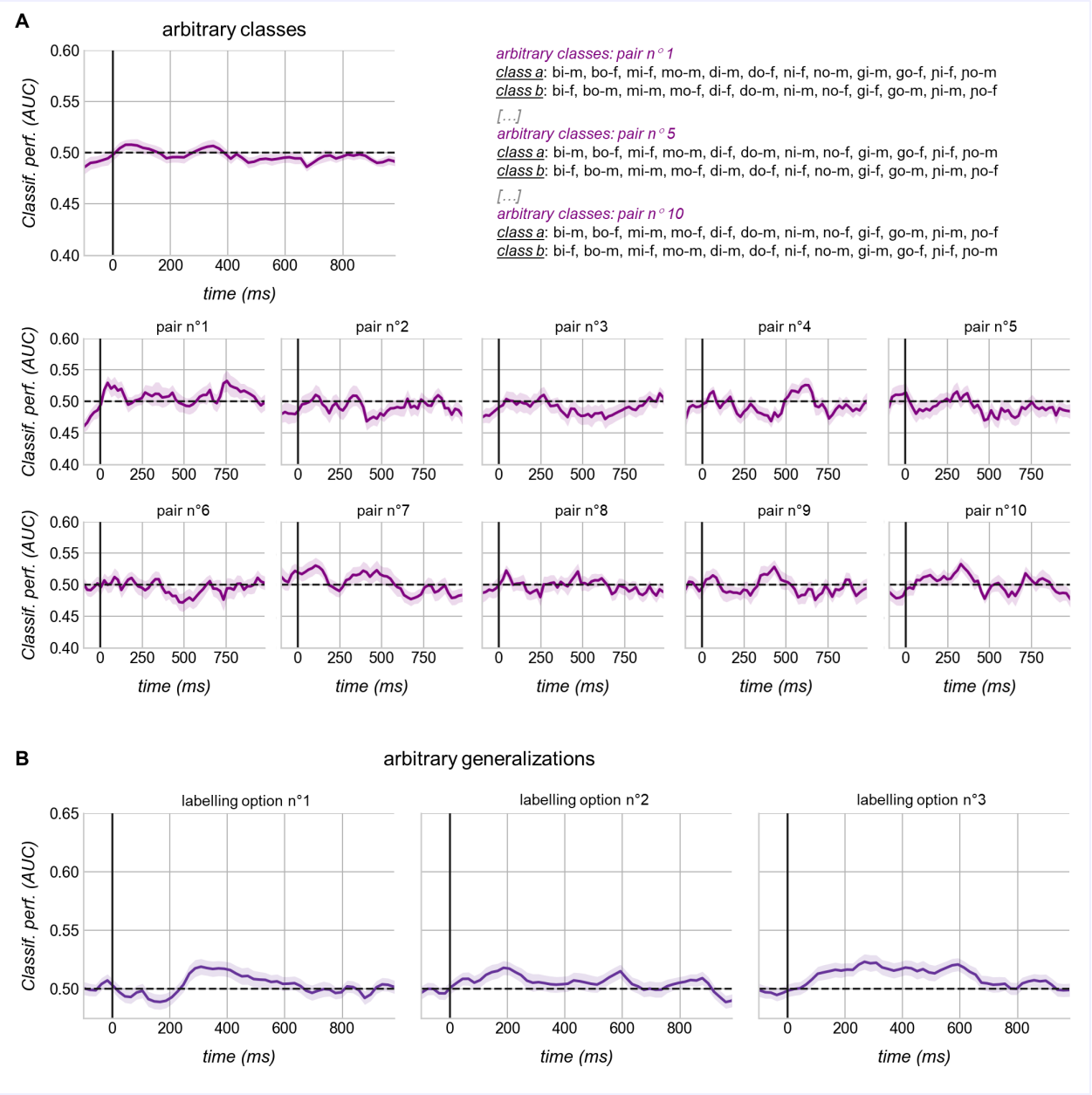

Figure 2.8 Sanity checks on classifier behavior and its interpretability.
As in all main decoding analyses, set of classifiers are trained and tested every 20ms along the ERPs. (A) The 24 sub-conditions composing the stimuli set (Figure 2.1) were partitioned in two arbitrary classes balanced in terms of consonant identities, vowel identities, speakers and their combination alternatives (i.e. consonant/vowel, vowel/speaker and consonant/speaker idiosyncratic pairings; examples are reported on the top right). Decoders were trained and tested on all possible pair of classes constructible with such partitioning. Note that we ensured the same kind of balance across classes (with the obvious exception of that concerning the investigated dimension) in all decoding problems presented in the main text. 2-tailed cluster-based permutations t-tests were used to detect any deviation from chance: none was found for the average performance across arbitrary pairs (top-left) nor for any single arbitrary contrast (bottom). Other

than the employment of a well-controlled stimuli set on our behalf, the absence of above/below-chance decoding confirms the reality of phonemes as psychophysical perception objects for infants. (B) This panel is the counterpart of Figure 2.4A, displaying arbitrary generalizations instead of those based on phonetic theory. Estimators trained on manner contrasts within each place context (e.g. /b/ vs /m/; Figure 2.4A-left) were tested upon place contrasts (e.g. /d/ vs /g/ and /n/ vs /ɲ/; as opposed to testing their ability to classify obstruents vs sonorants at alternative place contexts). To mimic Figure 2.4A, the resulting performance is reported by averaging according to each possible arbitrary labelling option: alveolar →obstruent & velar →sonorant (left), velar →obstruent & labial →sonorant (middle), alveolar →obstruent & labial →sonorant (right). Note that the three remaining labelling possibilities (i.e. velar →obstruent & alveolar →sonorant, labial →obstruent & velar → sonorant, labial →obstruent & alveolar →sonorant) produce the same outcomes, only reversed relative to chance. Cluster-based permutation t-tests (with the same settings used for the main analyses) revealed no above-chance scores, validating the traditionally-defined phonetic domains of manner and place of articulation as meaningful decompositional axes for the brain.
In both panels shaded areas correspond to the standard error (SEM) across subjects and dotted black lines mark theoretical chance level. The absence of filled circles indicates that no significant effect was found.

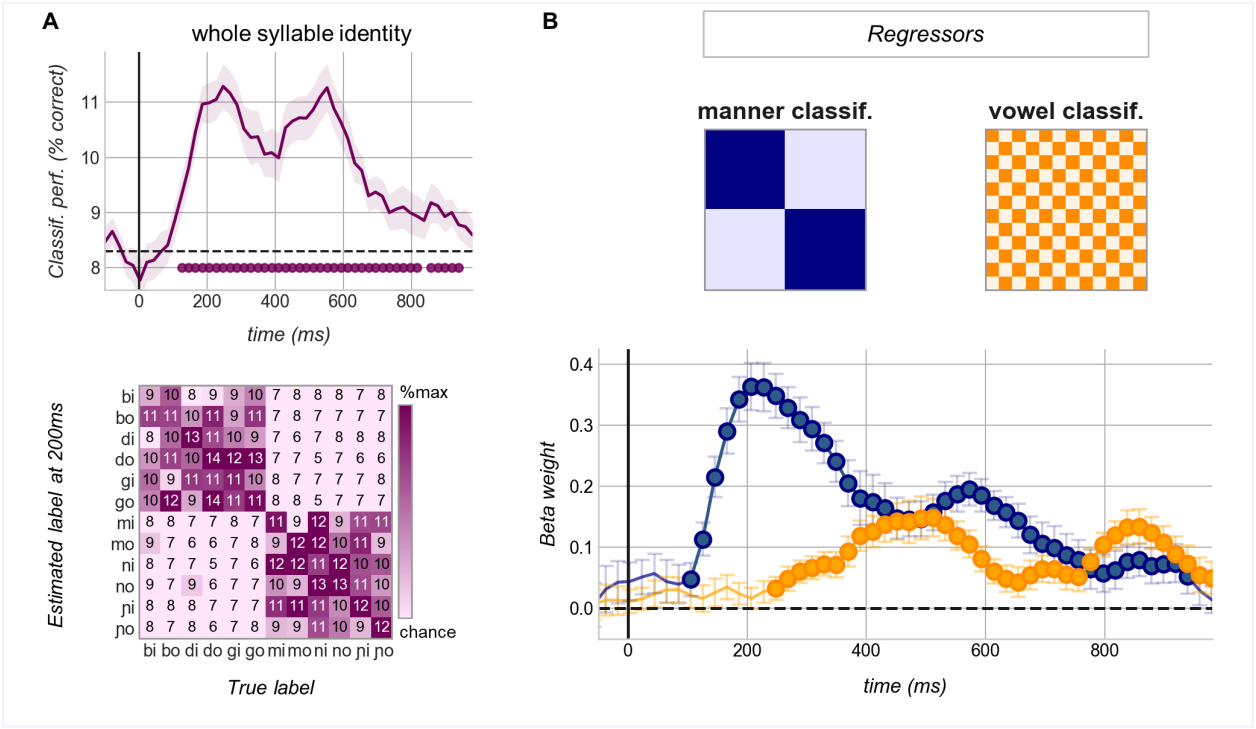

Figure 2.9 (complement of Figure 2.4).

(A) Top: time-resolved class-wise accuracy of estimators trained on syllable identities: “bi” vs “bo” vs ”di” vs ”do” vs ”gi” vs “go” vs “mi” vs “mo” vs ”ni” vs ”no” vs ” ɲi” vs “ɲo”. The shaded area corresponds to the SEM across subject, dotted black lines mark theoretical chance level, filled circles indicate when performance is significantly above chance (starting from 120ms: p_{clust}=0.0001) Bottom: confusion matrix yielded by the same classifiers at 200ms after stimulus onset. Numbers within each cell indicate the percentage of times a given syllable indicated along the x-axis was classified with the label reported on the y-axis. Off-diagonal values diverging from 0 signal misidentification (chance=8.3%). (B) Top: theoretical confusion matrices depicting a perfect separation between (i.e. the ideal classification of) manners of articulation and co-articulated vowel (classes are ordered as in A). Darker colors correspond to the values 16.6%; light colors correspond to 0%. These matrices were entered as predictors in the multiple regression analysis together with those illustrated by Figure 2.4C. Bottom: the obtained beta-weights, averaged across subjects and marked by filled circles when significantly above zero (100-920ms: p_{clust}=0.0001 for manner; 260-920ms: p_{clust}=0.0001 for the vowel). Vertical lines correspond to the SEM. Consistently with Figure 2.4A, this pattern of beta-weights shows that neural confusability was prominently driven by manner distinctions at first, but to a lesser extent later in the trial.

2.6. References

Arsenault, J. S., & Buchsbaum, B. R. (2015). Distributed Neural Representations of Phonological Features during Speech Perception. *Journal of Neuroscience*, *35*(2), 634–642. https://doi.org/10.1523/JNEUROSCI.2454-14.2015

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S. E., & Mehler, J. (1987). Discrimination in neonates of very short CVs. *The Journal of the Acoustical Society of America*, *82*(1), 31–37. https://doi.org/10.1121/1.395570

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General*, *117*(1), 21–33. https://doi.org/10.1037/0096-3445.117.1.21

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*(4), 711–721. https://doi.org/10.1037/0012-1649.29.4.711

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer* (Version 6.0.25) [Computer software]. http://www.praat.org/

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F. (2008). Hearing Faces: How the Infant Brain Matches the Face It Sees with the Speech It Hears. *Journal of Cognitive Neuroscience*, *21*(5), 905–921. https://doi.org/10.1162/jocn.2009.21076

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432. https://doi.org/10.1038/nn.2641

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906. https://doi.org/10.1121/1.1945807

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229. https://doi.org/10.1006/jpho.1999.0094

Choi, D., Dehaene-Lambertz, G., Peña, M., & Werker, J. F. (2021). Neural indicators of articulator-specific sensorimotor influences on infant speech perception. *Proceedings of the National Academy of Sciences*, *118*(20). https://doi.org/10.1073/pnas.2025043118

Choi, D., Kandhadai, P., Danielson, D. K., Bruderer, A. G., & Werker, J. F. (2017). Does early motor development contribute to speech perception? *The Behavioral and Brain Sciences*, *40*, e388. https://doi.org/10.1017/S0140525X16001308

Correia, J. M., Jansma, B. M. B., & Bonte, M. (2015). Decoding Articulatory Features from fMRI Responses in Dorsal Speech Regions. *Journal of Neuroscience*, *35*(45), 15015–15025. https://doi.org/10.1523/JNEUROSCI.0977-15.2015

Cummings, A., Madden, J., & Hefta, K. (2017). Converging evidence for [coronal] underspecification in English-speaking adults. *Journal of Neurolinguistics*, *44*, 147–162. https://doi.org/10.1016/j.jneuroling.2017.05.003

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, & R. van de Vijver (Eds.), *Variation and Gradience in Phonetics and Phonology* (pp. 265–310). Mouton de Gruyter.

Dehaene-Lambertz, G., & Pena, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport*, *12*(14), 3155–3158. https://doi.org/10.1097/00001756-200110080-00034

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, *22*(2), 109–122. https://doi.org/10.3758/BF03198744

Eimas, P. D., & Miller, J. L. (1980a). Discrimination of information for manner of articulation. *Infant Behavior and Development*, *3*, 367–375. https://doi.org/10.1016/S0163-6383(80)80044-0

Eimas, P. D., & Miller, J. L. (1980b). Contextual effects in infant speech perception. *Science*, *209*(4461), 1140–1141. https://doi.org/10.1126/science.7403875

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech Perception in Infants. *Science*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*(Aug), 1871–1874.

Fennell, C. T., & Waxman, S. R. (2010). What Paradox? Referential Cues Allow for Infant Use of Phonetic Detail in Word Learning. *Child Development*, *81*(5), 1376–1383. https://doi.org/10.1111/j.1467-8624.2010.01479.x

Fló, A., Brusini, P., Macagno, F., Nespor, M., Mehler, J., & Ferry, A. L. (2019). Newborns are sensitive to multiple cues for word segmentation in continuous speech. *Developmental Science*, *22*(4), e12802. https://doi.org/10.1111/desc.12802

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science*, *322*(5903), 970–973. https://doi.org/10.1126/science.1164318

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics*, *55*(6), 597–610. https://doi.org/10.3758/BF03211675

Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics*, *54*(3), 287–295. https://doi.org/10.3758/BF03205263

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

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

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2016). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Grosjean, F. (1996). Gating. *Language and Cognitive Processes*, *11*(6), 597–604. https://doi.org/10.1080/016909696386999

Halle, M. (2013). *From memory to speech and back: Papers on Phonetics and Phonology 1954-2002.* Walter de Gruyter.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, *180*, 4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005

Hillenbrand, James. (1983). Perceptual Organization of Speech Sounds by Infants. *Journal of Speech, Language, and Hearing Research*, *26*(2), 268–282. https://doi.org/10.1044/jshr.2602.268

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. https://doi.org/10.1111/j.1467-7687.2011.01089.x

Jain, A. K., & Chandrasekaran, B. (1982). 39 Dimensionality and sample size considerations in pattern recognition practice. In *Handbook of Statistics* (Vol. 2, pp. 835–855). Elsevier. https://doi.org/10.1016/S0169-7161(82)02042-2

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417–429. https://doi.org/10.1016/j.neuroimage.2017.06.030

Jusczyk, P. W. (2000). Early Research on Speech Perception. In *The discovery of spoken language* (pp. 43–71). MIT Press.

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*(5), 648–654. https://doi.org/10.1037/0012-1649.23.5.648

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. https://doi.org/10.1006/jmla.1993.1022

Jusczyl, P. W., Pisoni, D. B., & Mullennix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, *43*(3), 253–291. https://doi.org/10.1016/0010-0277(92)90014-9

Kanal, L., & Chandrasekaran, B. (1971). On dimensionality and sample size in statistical pattern classification. *Pattern Recognition*, *3*(3), 225–234. https://doi.org/10.1016/0031-3203(71)90013-6

Kazanina, N., Bowers, J. S., & Idsardi, W. (2018). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*, *25*(2), 560–585. https://doi.org/10.3758/s13423-017-1362-0

Khalighinejad, B., Cruzatto da Silva, G., & Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *The Journal of Neuroscience*, *37*(8), 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

King, J.-R., Pescetelli, N., & Dehaene, S. (2016). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. *Neuron*, *92*(5), 1122–1134. https://doi.org/10.1016/j.neuron.2016.10.051

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage*, *56*(2), 411–421. https://doi.org/10.1016/j.neuroimage.2011.01.061

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*(4), 649–662. https://doi.org/10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167–179. https://doi.org/10.1016/j.conb.2019.04.002

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J.-F. L., & Imada, T. (2014). Infants' brain responses to speech suggest Analysis by Synthesis. *Proceedings of the National Academy of Sciences*, *111*(31), 11238–11245. https://doi.org/10.1073/pnas.1410963111

Ledoit, O., & Wolf, M. (2003). *Honey, I Shrunk the Sample Covariance Matrix* (SSRN Scholarly Paper ID 433840). Social Science Research Network. https://papers.ssrn.com/abstract=433840

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, *68*(8), 1–13. https://doi.org/10.1037/h0093673

Lisker, L., & Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *WORD*, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Mahmoudzadeh, M., Dehaene-Lambertz, G., Fournier, M., Kongolo, G., Goudjil, S., Dubois, J., Grebe, R., & Wallois, F. (2013). Syllabic discrimination in premature human infants prior to complete

formation of cortical layers. *Proceedings of the National Academy of Sciences*, *110*(12), 4846–4851. https://doi.org/10.1073/pnas.1212220110

Mahmoudzadeh, M., Wallois, F., Kongolo, G., Goudjil, S., & Dehaene-Lambertz, G. (2016). Functional Maps at the Onset of Auditory Inputs in Very Early Preterm Human Neonates. *Cerebral Cortex*, bhw103. https://doi.org/10.1093/cercor/bhw103

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*(2 Part 1), 209–220.

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

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*(2), 91–121. https://doi.org/10.1016/S0010-0277(00)00109-8

Meert, W., & Van Craenendonck, T. (2018). *Time series distances: Dynamic Time Warping (DTW)*. Zenodo. https://doi.org/10.5281/zenodo.3276100

Mersad, K., & Dehaene-Lambertz, G. (2016). Electrophysiological evidence of phonetic normalization across coarticulation in infants. *Developmental Science*, *19*(5), 710–722. https://doi.org/10.1111/desc.12325

Mersad, K., Kabdebon, C., & Dehaene-Lambertz, G. (2021). Explicit access to phonetic representations in 3-month-old infants. *Cognition*, 104613. https://doi.org/10.1016/j.cognition.2021.104613

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

Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. *Cognition*, *13*(2), 135–165. https://doi.org/10.1016/0010-0277(83)90020-3

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229–240. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Nespor, M., Peña, M., & Mehler, J. (2003). On the Different Roles of Vowels and Consonants in Speech Processing and Language Acquisition. *Lingue e Linguaggio*, *2/2003*. https://doi.org/10.1418/10879

Ng, A. Y. (2004). Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. *Proceedings of the Twenty-First International Conference on Machine Learning*, 78. https://doi.org/10.1145/1015330.1015435

Nossair, Z. B., & Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *The Journal of the Acoustical Society of America*, *89*(6), 2978–2991. https://doi.org/10.1121/1.400735

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, *5*(11), eaay6279. https://doi.org/10.1126/sciadv.aay6279

Oller, D. K., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Lee, C.-C., Bowman, D. D., Long, H. L., Buder, E. H., & Vohr, B. (2019). Preterm and full term infant vocalization and the origin of language. *Scientific Reports*, *9*(1), 14734. https://doi.org/10.1038/s41598-019-51352-0

Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: Combining insights from microscopic and mass signals. *Trends in Cognitive Sciences*, *19*(3), 162–172. https://doi.org/10.1016/j.tics.2015.01.002

Park, A. S., & Glass, J. R. (2008). Unsupervised Pattern Discovery in Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 186–197. https://doi.org/10.1109/TASL.2007.909282

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, *25*(2), 143–170. https://doi.org/10.1016/j.newideapsych.2007.02.001

Port, R. F. (2010). Rich memory and distributed phonology. *Language Sciences*, *32*(1), 43–55. https://doi.org/10.1016/j.langsci.2009.06.001

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*, 130–150. https://doi.org/10.1016/j.cognition.2017.11.003

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494. https://doi.org/10.1037/0012-1649.39.3.484

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354. https://doi.org/10.1016/j.jneuroling.2009.12.004

Seidl, A., Cristià, A., Bernard, A., & Onishi, K. H. (2009). Allophonic and Phonemic Contrasts in Infants' Learning of Sound Patterns. *Language Learning and Development*, *5*(3), 191–202. https://doi.org/10.1080/15475440902754326

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Sinnott, J. M., & Gilmore, C. S. (2004). Perception of place-of-articulation information in natural speech by monkeys versus humans. *Perception & Psychophysics*, *66*(8), 1341–1350. https://doi.org/10.3758/BF03195002

Smits, R., Bosch, L. ten, & Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluation. *The Journal of the Acoustical Society of America*, *100*(6), 3865–3881. https://doi.org/10.1121/1.417242

Stemberger, J. P., & Stoel-gammon, C. (1991). THE UNDERSPECIFICATION OF CORONALS: EVIDENCE FROM LANGUAGE ACQUISITION AND PERFORMANCE ERRORS. In C. Paradis & J.-F. Prunet (Eds.), *The Special Status of Coronals: Internal and External Evidence* (pp. 181–199). Academic Press. https://doi.org/10.1016/B978-0-12-544966-3.50015-4

Stevens, K. N. (2000). *Acoustic Phonetics*. MIT Press.

Stokes, M. G., Wolff, M. J., & Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. *Trends in Cognitive Sciences*, *19*(11), 636–638. https://doi.org/10.1016/j.tics.2015.08.016

Sundara, M., Ngon, C., Skoruppa, K., Feldman, N. H., Onario, G. M., Morgan, J. L., & Peperkamp, S. (2018). Young infants' discrimination of subtle phonetic contrasts. *Cognition*, *178*, 57–66. https://doi.org/10.1016/j.cognition.2018.05.009

Swingley, D., & Aslin, R. N. (2002). Lexical Neighborhoods and the Word-Form Representations of 14-Month-Olds. *Psychological Science*, *13*(5), 480–484. https://doi.org/10.1111/1467-9280.00485

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding Words and Rules in a Speech Stream: Functional Differences Between Vowels and Consonants. *Psychological Science*, *19*(2), 137–144. https://doi.org/10.1111/j.1467-9280.2008.02059.x

Tsuji, S., Mazuka, R., Cristia, A., & Fikkert, P. (2015). Even at 4 months, a labial is a good enough coronal, but not vice versa. *Cognition*, *134*, 252–256. https://doi.org/10.1016/j.cognition.2014.10.009

van der Stelt, J. M., & Koopmans-van Beinum, F. J. (1986). The Onset of Babbling Related to Gross Motor Development. In B. Lindblom & R. Zetterström (Eds.), *Precursors of Early Speech: Proceedings of an International Symposium held at The Wenner-Gren Center, Stockholm, September 19–22, 1984* (pp. 163–173). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-08023-6_12

Vilain, A., Dole, M., Lœvenbruck, H., Pascalis, O., & Schwartz, J.-L. (2019). The role of production abilities in the perception of consonant category in infants. *Developmental Science*, *22*(6), e12830. https://doi.org/10.1111/desc.12830

Wang, K., & Shamma, S. (1994). Self-normalization and noise-robustness in early auditory representations. *IEEE Transactions on Speech and Audio Processing*, *2*(3), 421–435. https://doi.org/10.1109/89.294356

Westermann, G., & Reck Miranda, E. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, *89*(2), 393–400. https://doi.org/10.1016/S0093-934X(03)00345-6

Yang, X., Wang, K., & Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, *38*(2), 824–839. https://doi.org/10.1109/18.119739

Zhang, Q., Hu, X., Luo, H., Li, J., Zhang, X., & Zhang, B. (2016). Deciphering phonemes from syllables in blood oxygenation level-dependent signals in human superior temporal gyrus. *European Journal of Neuroscience*, *43*(6), 773–781. https://doi.org/10.1111/ejn.13164

Chapter 3. SPONTANEOUS ENCODING OF NUMBER BY THE INFANT BRAIN

ABSTRACT

The ability to handle discrete quantities permeates every aspect of every-day life. Recent neuroimaging investigations revealed that the adult brain regards approximate number as a primary attribute of the visual scene: numerosity is processed in complete automaticity by means of dedicated neural mechanisms. What is the ontogeny of these numerical intuitions? A large body of developmental studies demonstrated the availability of quantifications skills early in life, whose nature remains however unclear. Whereas classical views conceive the existence of an innate module specialized for numerical processing, various proposals point out how our numerical competence might be rooted in generalized and holistic estimates of size or intensity. To clarify this debate, we exposed 3-month-old infants to sets of either 4 or 12 items and recorded their neural responses with high-density electroencephalography (EEG). Our experiment entailed a strategic combination of carefully calibrated stimulus features and multivariate analytic techniques, enabling a reliable disentanglement of numerical and non-numerical effects. We observed that the infant brain encodes the (approximate) numerosity of auditory sequences in full automaticity (during sleep) and separately from concurrent quantities. Strikingly, estimators trained on the neural responses to sounds could successfully decode numerosity from the activity patterns elicited by visual arrays, uncovering a mechanism for the extraction of numerical information that transcends sensory modality, presentation format and arousal state. In a nutshell, our study provides neural evidence for a primitive and abstract number sense in humans.

3.1. INTRODUCTION

We are surrounded by magnitudes of all sorts: lengths and areas, weights and volumes, luminance, loudness, frequencies and durations. Among all possible magnitudes, number is a distinctive dimension in that is used to quantify collections of discrete objects. As literate adults, we are used to think of numbers in association with high-level constructs such as symbols, formulas and bank accounts. Yet, not only educated adults but also illiterates and even primates appear to rely on numerical estimations in their everyday life, to take decisions and to navigate the world (Gordon, 2004; Pica, 2004; Piantadosi & Cantlon, 2017). Considering the ubiquity of behaviors based on discrete quantity shall we ask: what is a number at its core?

Irrespective of their cultural background, all human adults can effortlessly approximate the cardinality of a set (i.e. its *numerosity*) without counting. Psychophysical assessments have shown that such approximate numerical judgments are susceptible to adaptation: repetitive exposure to a large group of items decreases their apparent number, and, vice-versa, repetitive exposure to small sets increases apparent numerosity (Burr & Ross, 2008). It is intriguing to highlight, in this regard, that adaptation is a perceptual phenomenon typically observed for basic sensory properties such as color, contrast or speed. Importantly, adaptation to numerosity occurs irrespective of other concomitant quantitative parameters such as the size, frequency and density of the stimuli (Burr & Ross, 2008).

In the last few years, neuroimaging studies have revealed that the visual system of adults extracts numerical information from the visual scene in a rapid, direct, and automatic fashion. That is, approximate numerosity is encoded at early stages along the cortical pathway (in extrastriate regions) and independently from concurrent quantitative dimensions. Further, such an extraction of purely numerical information is spontaneous and pre-attentive (Park et al., 2016; Fornaciai et al., 2017; DeWind et al., 2019; Lucero et al., 2020; Georges et al., 2020; Van Rinsveld et al., 2020). Just to mention an illustrative example from the results just summarized, Lucero and colleagues (2020) have employed backward masking in order to disrupt reentrant feedback from fronto-parietal areas to visual cortex. Strikingly, they have found that approximate numerosity is computed over occipital regions even in the absence of conscious awareness.

Taken together, behavioral and neural evidence highlight how the adult brain regards numerosity as a primary property of the visual scene, encoded irrespective of its relevance and by means of a dedicated process.

What are the developmental origins of this mechanism? Classical accounts have proposed the existence of a built-in system that is specifically responsible for approximate, non-symbolic and non-verbal numerical representations and fundamentally separate from other systems of magnitude processing. These accounts postulate an ontogenetic continuity, whereby infants would be able to approximate numbers, independently from the other quantitative variables, through the same system since birth (Dehaene, 1997; Feigenson et al., 2004; Spelke & Kinzler, 2007). Such a proposal

arises from the observation that what could be interpreted as adult-like signs of numerical processing appear very early in life. First, albeit far less precise, infant quantitative behavior mirrors that of adults in a characteristic aspect: it conforms to Weber's law. Specifically, both the capability of 6-month-olds to detect a numerical difference (Xu & Spelke, 2000; Xu et al., 2005) and their degree of preference for numerical changes over homogeneity (Libertus & Brannon, 2010) depend on the ratio between the presented numbers rather than their absolute difference. A second resemblance between early and mature quantitative processing was suggested by a handful of pioneer studies in the developmental neuroimaging field. While regions of the parietal cortex along the intraparietal sulcus (IPS) are widely known to play a crucial role in adult numerical cognition (for a concise review: Cantlon et al., 2009), an EEG investigation on 3-month-olds (Izard et al., 2008) and two fNIRs studies on 6-month-olds have reported that numerical changes (relatively to shape changes) trigger specific activity over right parietal areas (Hyde et al., 2010; Edwards et al., 2016).

Yet, although intriguing at first glance, a careful examination reveals how these parallels are poorly informative in the close. To start with, whereas a wealth of behavioral studies have reported that infant can readily distinguish between arrays of objects differing in their cardinality, a recent quantitative meta-analysis concludes that the evidential value of the data currently available does not support strong inferences (Smyth & Ansari, 2020). Most importantly, modern debates have emphasized the challenges inherent the exploration of numerical cognition, leading to a newly grown awareness on subtle confounds and possible misinterpretations that have been often overlooked. The latter arise from the fact that changes in number are necessarily accompanied by changes in a variety of other non-numerical parameters: to control for all of them at once is physically impossible (Leibovich & Henik, 2013). Despite the precautions taken by the investigators and the employment of clever and elegant ruses (e.g. the strategy developed by Xu & Spelke, 2000), a thorough qualitative review of the literature suggests that for most studies reporting numerical discrimination by infants it is always possible to identify non-numerical variables that were actually confounded with number (Mix et al., 2002; Rousselle et al., 2004; Cantrell & Smith, 2013). This leaves open the possibility that infants did not detect numerosity changes per se; instead, to compare groups of objects, they might have used one or a host of different magnitudes. The occurrence of such an eventuality is consistent with investigations reporting that numerical judgments of young children are highly susceptible to irrelevant perceptual features (Soltész et al., 2010): when comparing sets of objects the performance of 3-year-olds falls at chance precisely when non-numerical quantities are strictly controlled for (Rousselle et al., 2004). Further, it has been recently suggested that, when their discriminability is balanced, area is more salient than number during childhood while the reverse saliency pattern typical of adults might derive from formal education (Aulet & Lourenco, 2021a, 2021b).
Considering that any stimulus carries information about both numerosity and other quantitative dimensions, the parallelisms found between preverbal infants and adults at the neural level may need to be taken from an alternative perspective. Other than playing a pivotal role in adult numerical cognition, several studies have demonstrated how the posterior parietal cortex is involved in the

processing of other non-numerical quantities such as spatial length (Borghesani et al., 2019) and object size (Harvey et al., 2015). The neural representations of the letter are spatially intermingled with number-specific ones; whereas sophisticated imaging techniques (high-field 7T fMRI), allowing a much finer resolution than that achieved in infants studies, seem necessary to disentangle one kind of quantitative code from the other. As for what concerns the encoding of non-symbolic and approximate numerical information, parietal regions might be recruited only within certain circumstances: when information needs to be integrated over space but not over time (Cavdaroglu & Knops, 2019), when an active comparison is required (Cavdaroglu et al., 2015) or when attentional re-orienting is triggered by the experimental manipulation (DeWind et al., 2019; Lucero et al., 2020). Given these observations, the posterior parietal cortex might not hold a primary role in respect of the core and approximate numerical intuitions described above. It follows that the sensitivity of parietal regions to quantitative changes found early in development is undoubtedly meaningful but does not inform us upon the origins of a neural mechanism specialized for numerosity. Speaking of, a recent study based on steady-state visual evoked potentials (SSVEP) and a careful orthogonalization of quantitative parameters reports that selective neural sensitivity for approximate number is absent in 3-year-old children while it increases as a function of age (Park, 2018).

To recapitulate, the developmental studies conducted so far have uncovered the availability of quantitative mechanisms early in life, whose nature remains however unclear. In recent years, multiple theoretical accounts have proposed the existence of a generalized magnitude system that extends across various dimensions and that would correspond to the basic and unique quantitative system available at birth (Hamamouche & Cordes, 2019; Leibovich et al., 2017; Newcombe et al., 2015; Walsh, 2003). According to such proposals, young infants would encode and represent different kinds of magnitude through a common, undifferentiated code corresponding to a general "size" or "amount". Key to these views is the idea that purely numerical information is not readily trackable early on: we would acquire a specialized, dedicated sense for numerosity over the course of development. As a matter of fact, a general holistic code for quantity/magnitude would satisfactorily account for all the experimental observations mentioned so far and for additional findings. At 6 months the discriminability of numbers, durations and spatial extents is defined by the same Weber fractions[13] (VanMarle & Wynn, 2006; Brannon et al., 2006; Feigenson, 2007). The latter decline at the same speed over the course of the first year (Brannon et al., 2007), implying that to be developing might be a unitary system. Moreover, it has been shown that infants tend to create spontaneous mappings between different quantitative dimensions (de Hevia & Spelke, 2010; Lourenco & Longo, 2010). For instance, 1 to 3 day-olds presented simultaneously with auditory numerical sequences and visual line lengths manifest magnitude-congruent expectations across

[13] Note that in adults Weber's law governs not only the discriminability of numerosities but also that of a wide variety of other dimensions (Cantlon et al., 2009). Therefore, the observation of size/distance effects corresponds to a vague piece of evidence in itself.

number, time (duration) and space, indicating, at the very least, an innate sensitivity to their common structure (de Hevia et al., 2014). Overlapping developmental trajectories and cross-dimensional transfer might be indicative of a generalized sense of magnitude available since birth. Still, they remain equally compatible with the possibility that different types of quantity are encoded in distinct, dedicated formats and compared or processed by means of common mechanisms only downstream.

The present study

Given the widespread idea that approximate numerical computations function as a "start-up tool" for the acquisition of mathematics (Piazza, 2010), and the consequences this idea has for educational and rehabilitative interventions (Butterworth, 2018), a deeper understanding of early human quantification appears crucial. In the current study, we investigated the existence and characteristics of a core neural system specifically dedicated to numerical processing. If the human brain regards numerosity as a primary perceptual descriptor since start, we expected very young infants to encode numerical information just as adults: automatically and pre-attentively, separately from concurrent non-numerical parameters.

To address this possibility, we recorded high-resolution event-related potentials (ERPs) from 3-mo-old infants while they were exposed to sequences of either 4 or 12 naturally rich orchestral string tones (24 different auditory sequences were played in a randomized order for a total of ~ 1900 trials /subject). Unlike any previous developmental paradigm in this domain, our auditory stimulation did not probe any change detection mechanism, did not involve any comparison process and did not require active attendance. On the contrary, participants were most often asleep, enabling the assessment of completely automatic neural processes.

Relatively to the typical employment of visual displays, where the relationship among sensory cues is unavoidably non-linear (Mix et al., 2002; Lipton & Spelke, 2004), auditory arrays allowed a straightforward traceability of all the quantitative dimensions involved. In order to ensure the latter, the auditory space (Figure 3.1A) was constructed such that each pair of numerical conditions equal in duration ("12S" and "4M"; "12M" and "4L") incorporated the same amount of sound/silence and each pair of sequences equal in rate ("4M" and "12M"; "4L" and "12L") had also the same tone length

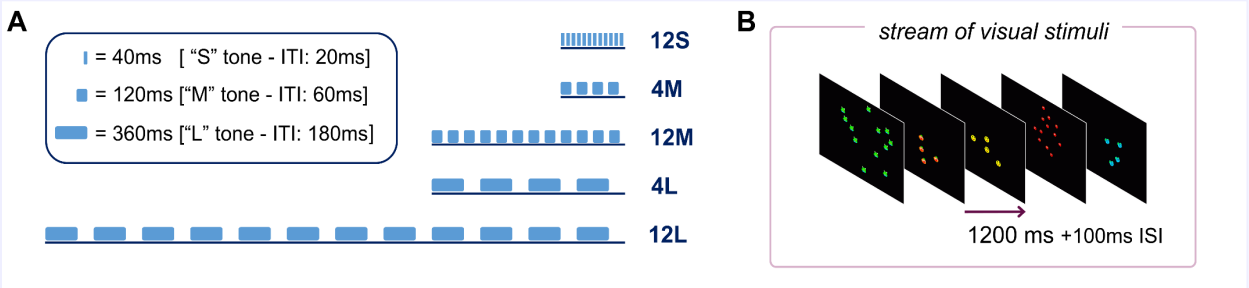

Figure 3.1 Experimental paradigm. (A) auditory sequences (ITI=inter-tone-interval); their non-numerical quantitative characteristics are reported in Table 1. (B) Visual stimulation (ISI= inter-stimulus-interval)

and inter-tone-intervals (Table 3.1). Inspired by recent efforts on adults, the combination of this design with strategic/tactical multivariate analyses enabled to isolate the effects of numerosity from that of the other quantities. Lastly, when possible (either at the beginning or at the end of the experimental session), infants were presented with the same numerosities in a radically different format: visual images containing (simultaneous) sets of either 4 or 12 colorful objects (Figure 3.1B).

Condition / *Parameters*	***4M***	***4L***	***12M***	***12S***	***12L***
tone duration (ms)	120	360	120	40	360
inter-tone-interval (ms)	60	180	60	20	180
total sequence duration (ms)	720	2160	2160	720	6480
tone rate (Hz)	5.6	1.9	5.6	17	1.9
total amount of sound (ms)	480	1440	1440	480	4320
total amount of silence (ms)	240	720	720	240	2160

Table 3.1: physical parameters characterizing each auditory condition. A given tone duration is always matched to a specific inter-tone-interval and their ratio is constant (2:1), leading to a one-to-one correspondence between tone duration and rate. As a consequence, when sequenced of 4 vs 12 sounds are matched for their total duration (e.g. 12M and 4L), they also embed the same amount of sound (e.g. 1440ms) and silence (e.g. 720ms). This design was crucial because it ensured the possibility to separate numerosity from any other possible quantitative parameter characterizing the stimulation. Any time our analysis controls for tone rate, tone duration and inter-tone-interval are also accounted for and any time our analysis accounts for sequence duration, the total amount of sound and the total amount of silence are also accounted for. This design does not enable to disentangle the effects of tone duration from that of inter-tone-interval and the effects of the overall sequence duration from those of the total amount of sound/silence: this kind of investigation was beyond the scope of the study.

3.2. MATERIALS & METHODS

3.2.1. Participants

26 normal-hearing infants (15 females and 11 males) were tested between 12 and 14 weeks after birth (mean age= 13 weeks and 2 days). All infants were full-term, with the exception of a female and a male participants, born at a gestational age of 37 weeks at a healthy weight (2800gr for the female

and 3250gr for the male, i.e. above the 10th percentile[14]). None of the participants had pre- or post-natal neurological issues or known genetic disorder. An additional 23 subjects were recruited but excluded from analysis because of: impossibility to collect enough data due to excessive fussiness during the experimental session (n=13), insufficient number of trials after artifact rejection (n=4, the artifact rejection procedure is described below), technical problems during data acquisition (n=6). The protocol was approved by the regional ethical committee for biomedical research (CPP Region Centre Ouest 1). Parents gave their written informed consent before starting the experiment. A certificate and a baby book were provided as thanks for the participation.

3.2.2. Stimuli

Orchestral string tones were synthesized with Ableton Live (software version 10; Berlin, Germany). They consisted of four notes (C_3, G_3, C_4 and G_4) played by either a viola or a cello and had three possible item durations (40, 120 and 360ms; referred to as "S", "M" and "L" respectively) for a total of 24 different sounds. Each auditory sequence presented a single tone repeated either 4 or 12 times at a constant rate. The duration of inter-tone intervals were half that of the single tone characterizing the sequence (i.e. 20ms for S sounds, 60ms for M sounds or 180ms for L sounds) resulting in three possible rates (1.9Hz, 5.6 Hz and 17Hz) and four possible total durations (240, 720, 2160 and 6480ms; however the shortest total duration, i.e. "4S", was excluded a-priori from the analysis). Crucially, the minimal proportional difference between numbers, tone rates or total sequence durations was always 1:3. Given that infant numerical and temporal discrimination depend on the same ratios (Brannon et al., 2007), this means that the three quantities were equally discriminable for our subjects. All sequences had the same sound-to-silence ratio (2:1) and equal loudness (75 dB). They were placed on the left audio channel, connected to the loudspeakers. Clicks were positioned on the right channel in correspondence to the onset of the initial and the offset of the final tone for each sequence; these clicks were used as a TTL signal to ensure a precise mapping between the EEG recording and the stimulation.

The visual stimuli were the same used by Izard, Dehaene-Lambertz & Dehaene (2008). They consisted of a set of 400 images depicting either 4 or 12 colorful animal-like objects on a black background. To minimize any possible effect ascribable to non-numerical perceptual attributes, the position of the objects and the physical parameters of the image varied across stimuli following two different rules. Namely, in 200 images the extensive parameters of the display (total luminance and total occupied area) were kept constant across numerosities, whereas in the remaining 200 images the intensive parameters (object surface size, average area devoted to each object) were equated

[14] as indicated by the World Health Organization https://www.who.int/tools/child-growth-standards

between numerosities. For more details on the control for non-numerical factors see Izard, Dehaene-Lambertz & Dehaene (2008).

3.2.3. Procedure

Infants were tested in a soundproof Faraday cage equipped with a computer screen and loudspeakers on the ceiling. They were held by a caregiver in a comfortable position and constantly monitored by the experimenter from two video cameras located underneath and above the screen. All stimuli were presented using the Python package PsychoPy (Peirce, 2007).

Auditory sequences were broadcast through the loudspeakers with an inter-sequence-interval (ISI) fixed at 1 second. The order of the stimuli was randomized with two constrains: the same numerosity could not be presented more than 4 times in a row and the same auditory sequence (characterized by a given note, instrument, numerosity, rate and duration) could not be repeated more than twice in a row. The auditory stimulation was organized in blocks of 688 sound sequences where notes, instruments and numerosities were perfectly balanced. Each block was composed of 72 "4S", 136 "4M"/"4L"/"12M" and 104 "12S"/"12L" sequences. A misbalance in experimental conditions was motivated by analytical plans (e.g. 4S was included in the stimulation to provide perceptual harmony but excluded from the analysis a-priori) and practical constrains (e.g. the need to collect a large number of trials in a reasonably limited amount of experimental time). A minimum of 2 blocks (corresponding to ~71 minutes of listening) and a maximum of 3 blocks were presented to each participant. Breaks were taken whenever necessary and sleep strongly encouraged. On average, subjects were asleep for 72% of the auditory session.

Participants were exposed to visual displays only when possible and appropriate on the basis of their psychophysiological state (e.g. awake and calm vs. sleepy or restless). Images were organized in mini-blocks of 100 items, where numerosities and non-numerical parameter control strategies were perfectly balanced. Images were presented for 1200ms in a continuous stream and randomized order, interspersed with 100ms-long blanks. The onset of each image was recorded through a photodiode capturing the appearance of a white rectangle at the bottom corner of the computer screen (not visible for the subject). When the infant looked away, the stream of numerical displays was interrupted by a colorful attractor until attention was re-established. The visual stimulation finished after the presentation of all the 400 images available or as soon as the participant could no longer engage with the displays. Out of the 26 subjects included in the final analysis, 13 infants attended to visual images before the auditory stimulation, 1 infant partly before and partly after the auditory sequences and 6 participants were presented with images only in the conclusive portion of the experimental session.

3.2.4. EEG recording and data preprocessing

The electroencephalogram (EEG) was continuously digitized at 500 Hz (Net Amps 300 EGI amplifier combined with NetStation 5.3 software) from 256 channels. We used a prototype HydroCel net (EGI; Eugene, OR, USA) referenced to the vertex featuring an intensive coverage of temporal areas (Figure 1.3).

The first preprocessing steps consisted in applying a band-pass filter ([0.5 - 40Hz]) and setting the mean voltage of each electrode to zero. We then followed an artifact detection-correction procedure similar to that used in Chapter 2. Namely, we based artifact detection on adaptive (rather than absolute) thresholds in order to account for inter-individual variability and the heterogeneous influence that reference distance and vigilance state exert on the voltage. We used series of algorithms that rejected samples on the basis of: the voltage amplitude and its first derivative; the variance across a 500ms-long moving time window; the fast running average and the deviation between the fast and the slow running averages within a 500ms-long sliding time window. Thresholds were set independently for each subject and for each electrode upon the distribution of these measures along the whole recording (threshold = median +/- n*IQ, where IQ is the interquartile range of the distribution). Two additional algorithms identified whether the power within the 0-10Hz band was excessively low or within 20-40Hz excessively high relative to the total power; and whether the voltage amplitude displayed by each sensor at a given time point was disproportionate relative to that recorded by the other sensors at the same instant. For these last two algorithms, thresholds were computed upon the distribution across channels.

We followed an iterative detection-correction procedure, where previously identified bad samples were kept aside for the subsequent artifact detection steps. At each run, the output of the artifact detection procedure consisted of a rejection matrix with the same size of the EEG data. We started by applying the detection algorithms (2 runs) to the continuous recording in order to identify very short signal disruptions (80ms max), corresponding to heart beats or jumps. We corrected these very short segments by estimating their principal components (PCA) and removing the first *n* components determining 90% of the variance. After using a high-pass filter (0.5Hz) to eliminate possible drifts created by such operation, we applied the detection algorithms twice more.
At this point the EEG data (and the corresponding rejection matrix) were segmented into epochs from -400ms to +1200ms relative to the offset of the last tone composing the auditory sequences and from -200ms to +1300ms relative to the onset of the images. A third set of epochs was crafted around the onset of the auditory sequences and included a different time-window depending on the condition: from -200ms to + 1820ms relative to the onset of the first note for “4M” and “12S” trials; -200 to +3260ms for “4L” and “12M” trials; -200ms to +3260ms and +2500ms to +6500ms for “12L” trials (i.e. the longest sequences were divided in two parts spanning from the onset of the first L tone until the end of the 6th inter-tone interval and from the onset of the onset of the 6th L tone until the 12th inter-tone interval).

We used the (segmented) rejection matrices to mark time points with prominent artifacts (*bad times*) and channels that did not function properly (*bad channels*). Specifically, *bad times* were periods longer than 80ms with a percentage of rejected channels superior to 30% or beyond 2IQ from the 3rd quartile of the distribution of the percentage of rejected channels across time. Similarly, *bad channels* were the ones not working properly for more than 30% of time or with a percentage of bad samples that went beyond 2IQ from the 3rd quartile of the distribution of the percentage of rejected samples across channels. . We corrected *bad channels* and long rejected segments that did not contain *bad times* using spherical splines interpolation (Perrin et al., 1989). However, spatial interpolation was carried out only if at least 50% of the neighboring channels were intact. Periods defined as *bad times* were not corrected because there was not enough information available to reconstruct the signal.

Trials were discarded if more than 15% of their samples contained artifacts or if more than 2.5% of their channels were marked as bad. Epochs were also discarded based on their Euclidean distance from the average, i.e. when their mean or maximum distance from the average response was an outlier in the distribution (> 3rdquartile + 1.5*IQ). Following automated rejection, the remaining epochs were visually inspected and a few channels or trials still presenting obvious aberrancies were dropped.
Since our paradigm was mainly based on the auditory stimulation, our inclusion criterion concerned the latter: participants were included in the study with a minimum of 192 artifact-free epochs for *each* of the most frequent auditory conditions ("4M", "4L", "12M"). In our final group of infants (N=26), the mean rejection rate for auditory trials was 33.5% (12.4 to 48.4%); on average, the number of artifact-free epochs available per subject was 247 for "4M"/"4L"/"12M" and 189 for "12S"/"12L". For visual trials to be included in the analysis, subjects were required to have at least 64 artifact-free epochs. The mean rejection rate for visual epochs was 62% (25 to 77.4%) and only 16 out the 20 subjects who attended the visual displays met the criterion. The average number of artifact-free visual epochs available for these 16 subjects was 109.

Before submitting them to the main analyses, epochs were low-pass filtered at 20Hz and mathematically re-referenced to the mean of all channels. Time-resolved multivariate pattern analyses were conducted within subject, relying on the Python packages MNE (Gramfort et al., 2013, 2014) and Scikit-Learn (Pedregosa et al., 2011).

3.2.5. Decoding

EEG data was first prepared by removing a linear trend from the entire segments, at the aim of reducing eventual slow drifts, and then dividing epochs into 110 consecutive windows of 10ms, from -40ms to 1060ms relative to the onset of the last tone composing the auditory sequence. All the procedures described in this section were carried at the level of single time-windows, each corresponding to a matrix with the shape *n* channels *x* 5 samples (sampling rate = 500Hz, 5 samples=10ms). The general goal of the decoding analyses was to predict a vector of binary

categorical data (*y*, containing the classes "4" vs "12") from a matrix of single-trial neural data (*X)* which included all EEG channels.

For the main analysis we used three separate sets of estimators and followed the three complementary strategies illustrated by Figure 3.2A. In each training phase, one class included trials belonging to a single experimental condition, while the alternative numerical class was composed of two experimental conditions, one characterized by the same tone rate and the other characterized by the same total duration found in the homogeneous class. With this design, numerosity was the one and only reliable feature for estimators to learn how to separate classes, being a specific tone rate/sequence duration distinctive for a given class only in certain cases. Further, to make sure that duration-based or rate-based learning could not lead to successful performance, each set of classifiers was tested twice (Figure 3.2A). In a first test, sequence duration could not drive above chance scores since it was the same across test conditions and the specific rate indicative of the composite class during training could not lead to above-chance scores since it was either not at all present (A1, B1) or a peculiarity of the opposite numerosity (thus misleading, C1). Within a second test, tone rate could not drive above chance scores since it was the same across test conditions and the total duration indicative of the composite class during training could not lead to above-chance scores since it was either absent (A2, C2) or a feature of the opposite numerosity (thus misleading, B2).

In order to avoid overfitting, we used a cross-validation procedure with 100 loops. Trials were shuffled at each run, then assigned to the respective training and test sets. Concerning schema A, the number of "4M" trials was first equated to that of "4L" (by randomly selecting and dropping *n* extra epochs for the most numerous condition), then 15% of "4M" trials and 15% of "4L" trials were kept aside for the test phase. The splitting was slightly different in schemas B and C in order to counterbalance the fact that "12S"/"12L" trials were less numerous relative to "12M" (see *Procedure* above). Namely, 80% and 20% of "12S"/"12L" trials were assigned to the training and test set respectively. The splitting of "12M" trials was then calibrated to obtain a balanced training set in terms of number of epochs per "12" condition (e.g. for schema B: n 12M trials in test set = total number of 12M trials available – n of "12L" trials in training set)[15]. This partitioning was performed in a stratified fashion such that all sources of irrelevant variability (i.e. musical notes and instruments) were distributed in equal proportions. When a specific condition was used only within training or exclusively at test, all the corresponding trials were assigned to one of the two sets according to the schema at hand.

Once established the training and the test set for a given run, we applied a "micro-averaging" procedure, a strategy commonly employed on adults to improve signal-to-noise ratio (Grootswagers

[15] To recapitulate, the splitting was always organized to ensure balanced composite classes (in terms of *n* epochs per condition) and still exploit all 4M/4L/12M trials available for a given subject.

et al., 2016). Within each experimental condition, this consisted in shuffling the epochs and then forming pseudo-trials by averaging together (randomly-defined) groups of 8. At the end of such operation, we balanced the test sets by equalizing the number of micro-averaged epochs across numerosity classes. In practice, we randomly selected the same amount of pseudo-trials available for the least numerous class from the most abundant.

Next, following the z-scoring each feature (i.e. channel and time point across trials), a L2-norm regularized Logistic Regression was fitted to the training set (Fan et al., 2008) in order to find the hyperplane that could maximally predict *y* from *X* while minimizing a loss function. Since composite classes contained more trials than heterogeneous ones, a weighting procedure was applied in order to equalize the contribution of each class to the definition of the hyperplane. The other model parameters were kept to their default values as provided by the Scikit-learn package.
After training, the models were used to predict *y* from the test set and their performance was evaluated by comparing estimates to the ground truth. All algorithms produced as an outcome vectors of probabilistic estimates. These probabilities were scored by computing the area under the Receiver Operating Characteristic curve (AUC), which summarizes the ratio between true and false positives. The value of AUC ranges between 0 and 1, with 0.5 corresponding to chance level. The scores obtained across loops and from either all or a group of train/test schemas were averaged within subject before submitting them to statistical analysis.

Within the same cross-validation loop, estimators were tested both at the trained time sample and on all the other 109 windows. The outcome of this procedure is a temporal generalization matrix (King & Dehaene, 2014) where each row reports the classification scores of a single estimator trained at time *t* and tested all along the trial (each time lag *t'* corresponds to one column). When a neural code is recursive or sustained, a successful estimator trained at a given time point (i.e. specific to a given pattern of brain activity) will achieve above-chance scores not only at the same time point but also at other time lags. Thus, the shape of the generalization performance within the temporal matrix can provide rich insights upon the dynamics of the neural activity patterns enabling classification.

Generalization across sensory modalities

In a second decoding analysis we investigated whether the infant brain processes the numerosity embedded in auditory and visual displays through a common neural code. We used the same pipeline as above (100 cross-validation loops, L2-regularized logistic regression with weighted class contribution etc.) but this time probed decoder's ability to predict *y* from the neural responses to visual displays. Given the divergence of training and test data, this analysis entailed the opportunity to employ all auditory conditions at once, with the potential benefit of increasing predictive power. Yet, in order to prevent class separation from being based on non-numerical parameters, it remained crucial to keep our three training schemas (Figure 3.2A) separate. Following such considerations, in order to maximize both sensitivity and specificity, we exploited an inherent property of the learning process: iterations. That is, the optimal hyperplane is computed though successive, intermediate and

approximate minimizations of the cost function, while the model is updated incrementally after each pass over the dataset. Building on this principle, we used a single set of decoders (one estimator for each of the time lags that led to successful classification in the main analysis) and trained them in an online fashion, following the same strategy as before. Specifically, the initial pipeline was modified such that the training set (i.e. schema) changed (randomly) at each internal iteration, for a total of 600 *partial fits.* The final weights of the model corresponded to the average value of the coefficients computed across all updates. Overall, this strategy enabled us to capitalize on the possibility to use a larger training set and still minimize the impact of non-numerical parameters on learning (adopting the same training logic of the main analysis).

Visual data was prepared for tests in the same way as auditory data. Before micro-averaging, we made sure to equalize the amount of trials controlled for extensive parameters to that of trials controlled for intensive properties within each numerosity condition. When the number of remaining trials was too scarce to obtain a minimum of 5 pseudo-trials/visual numerosity (4 subjects), we used some of the single epochs more than once, with the constraint that two pseudo-trials could not share more than 2 single epochs (out of 8).
As before, each classifiers trained on a given time-window *t* (in between 400 and 800ms after the onset of the last tones) was tested at every time-lag from 0 to 1000ms after the onset of the image (x-axis in Fig. 4A). Obtaining such temporal generalization matrix was essential for this analysis since we had no a-priori hypothesis concerning the temporal delay of numerical estimation within the visual modality.

Finally, to exclude the eventuality of non-numerical confounds on the observed performance, we created two supplementary test sets: one exclusively composed of trials with extensive parameter control (i.e. a group of trials in which object size and area devoted to each object co-varied with numerosity) and the other including only those trials controlled for intensive elements (i.e. trials in which total occupied area and luminance increased as a function of numerosity).

3.2.6. Representation Similarity Analysis

The aim of this analysis was to test whether numerical and non-numerical information could be dissociated from the activity patterns evoked by the auditory sequences. Crucially, unlike classification-based decoding, Representational Similarity Analysis (RSA) allows to assess the effect of multiple quantitative variables at once (Castaldi et al., 2019). The general outline of the analysis consisted in modelling a set of neural (i.e. empirical) dissimilarity matrices, one for each time point, as a linear combination of 3 theoretical matrices providing all together an exhaustive description of the quantitative information embedded in the auditory space.

To compute neural dissimilarity, we started by down-sampling the EEG recordings (with a moving average of 4 time points) to 125Hz, then averaged together the epochs belonging to the same condition. Given that the potential of this kind of analysis is best expressed with rich experimental

designs (Kriegeskorte, 2008), we averaged trials where notes were played by different instruments separately. That is, for each main condition (Figure 3.1A) we obtained two evoked responses, corresponding to the sub-conditions "viola" and "cello". Finally, we calculated the correlational distance (1-Pearson across channels) between each pair of sub-condition. To counterbalance the fact that "12S/12L" trials where less numerous (see *Procedure*), we repeated this computation 100 times. At each loop, the evoked responses were calculated by averaging an equal amount of trials per sub-condition (for each of sub-condition, we randomly selected the same amount of trials corresponding to that available for the least abundant). In this way, we made sure that each condition had the same signal-to-noise ratio and still exploited all the data available for a given subject, thereby optimizing the stability of the estimates. The final neural dissimilarities corresponded to the mean distances obtained across the 100 loops.
The theoretical dissimilarity matrices defined the distance, on a logarithmic scale, between each pair of sub-conditions along the quantitative dimensions defining the auditory sequences: number, tone rate and total sequence duration. The three matrices were entered as predictors in a linear multiple regression in order to explain the neural distances observed at each time point. All the dissimilarity matrices were z-scored before estimating the regression coefficients. As a result, for each subject, we obtained a set of beta weights reflecting the portion of the variance that each of the predictor matrices *uniquely* explained in the evoked activity patterns over time.

Concerning the RSA performed at sequence offset (Figure 3.3A), "12L" trials were excluded from the analysis in order to balance out the design. Thanks to such expedient, all three predictors remained adequately decorrelated (number-rate: 0.26, number-duration: -0.17, rate-duration: 0.26) and their variance inflation factors (VIF) satisfactorily low (1.143, 1.19, 1.143 for number, tone rate and total sequence duration respectively). Note that reducing multicollinearity at the minimum is important, given that strong correlations between predictors/high VIFs can compromise the reliability of the outcome coefficients.
For the RSAs performed within sequence (Figure 3.3B and 3.6), the analysis was restricted to "12" trials and no exclusion was needed in order to obtain balanced distance matrices. Starting from the epochs crafted around sequence onset (see *Data preprocessing*), we obtained two sets of evoked activity patterns, corresponding to two cardinalities, by cropping the signal from the onset of the N^{th} note and up to 800ms thereafter. For the main analysis we chose to contrast "3" and "7" in order to parallel the RSA at sequence offset in the best way possible, thereby obtaining a set of beta weights that could be interpreted in relation to the former. For this contrast, total sequence duration (in this case: the time elapsed from sequence onset up to the specific cardinality) was equal between "7S" and "3M" (i.e. 360ms) and between "7M" and "3L" (i.e. 1080ms), mirroring the correspondence in total duration of "12S"/"4M" and "12M"/"4L" (Figure 3.1A). As in the previous case, this characteristic contributed to keeping multicollinearity at the minimum (VIFs were 1.059, 1.565, 1.555 for number, rate and duration respectively). Further, selecting the cardinalities "3" and "7" enabled to (a) minimize onset-related low-level effects, (b) focus on a portion of the signal that was

sufficiently far from sequence offset not to overlap with the main RSA analysis (c) test a numerical ratio greater than 1:2, taking into account behavioral observations reporting that a 1:2 ratio is not sufficiently large for newborns to discriminate numerical displays (Izard et al., 2009).

3.3. Statistical analysis

To calculate statistics we performed second-level tests across subjects employing the MNE dedicated functions. Following a standard approach in adult studies (e.g. King et al., 2016), we used one-sample cluster-based permutation t-tests (Maris & Oostenveld, 2007) which intrinsically account for multiple comparisons. We tested whether (a) time-resolved classification scores were higher than chance and (b) whether multiple regression beta-weights differed from zero. The analyses considered two-dimensional clusters for decoding scores (i.e. they were always performed on the entire temporal generalization matrix) and one-dimensional clusters in the case of regression coefficients. Univariate t-values were calculated for every score/beta-weight with the exclusion of those corresponding to the baseline period. All samples exceeding the 95th quantile were then grouped into clusters based on temporal adjacency. Cluster-level test statistics corresponded to the sum of t-values within each cluster. Their significance was computed by means of the Monte-Carlo method: they were compared to a null distribution of test statistics created by drawing 10000 random sign flips of the observed outcomes. A cluster was considered as significant when its p-value was below 0.05.

3.4. RESULTS

In our main analysis, we asked whether (and when) linear classification algorithms could decode “4” vs “12” from infant fine-scale evoked activity patterns, irrespective of the particular tone rate and total duration characterizing the sequences that were played. With this aim, we used a strategic combination of training and test sets to ensure that successful performance could not be ascribable to non-numerical effects (Figure 3.2A). Specifically, within three distinct sessions, the algorithms were trained to separate one experimental condition from a composite class that included sequences matched in either rate (50% of cases) or duration. At test, the non-numerical quantities distinctive for one particular numerical class during training were prevented from leading to above-chance scores since they characterized either both numerosities (A1, A2, B1, C2), the opposite number (relatively to the training set: B2, C1), or none of the test trials. The three decoding schemas illustrated in Figure 3.2A were applied on brief (10ms) consecutive windows starting from the onset of the last tone composing the sequences and for 1 (silent) second thereafter. Figure 3.2B shows that all estimators trained in between 440 and 750ms achieved above-chance scores, with the best classification performance observed at 610ms after the onset of the last tones (N=26; M=0.557±0.044, chance=0.5). When systematically tested at the other time points along the trial (King & Dehaene, 2014), these estimators yielded similar cross-temporal classification dynamics:

their scores raised above chance level always around 400ms, peaked at ~600ms and fell at chance after 750ms (Figure 3.2C). Such square-shaped generalization reveals that the neural activity pattern underlying decodability consisted of an essentially stationary code. Further, Figure 3.2D shows that these classification dynamics were qualitatively similar across tests (see also Figure 3.5).

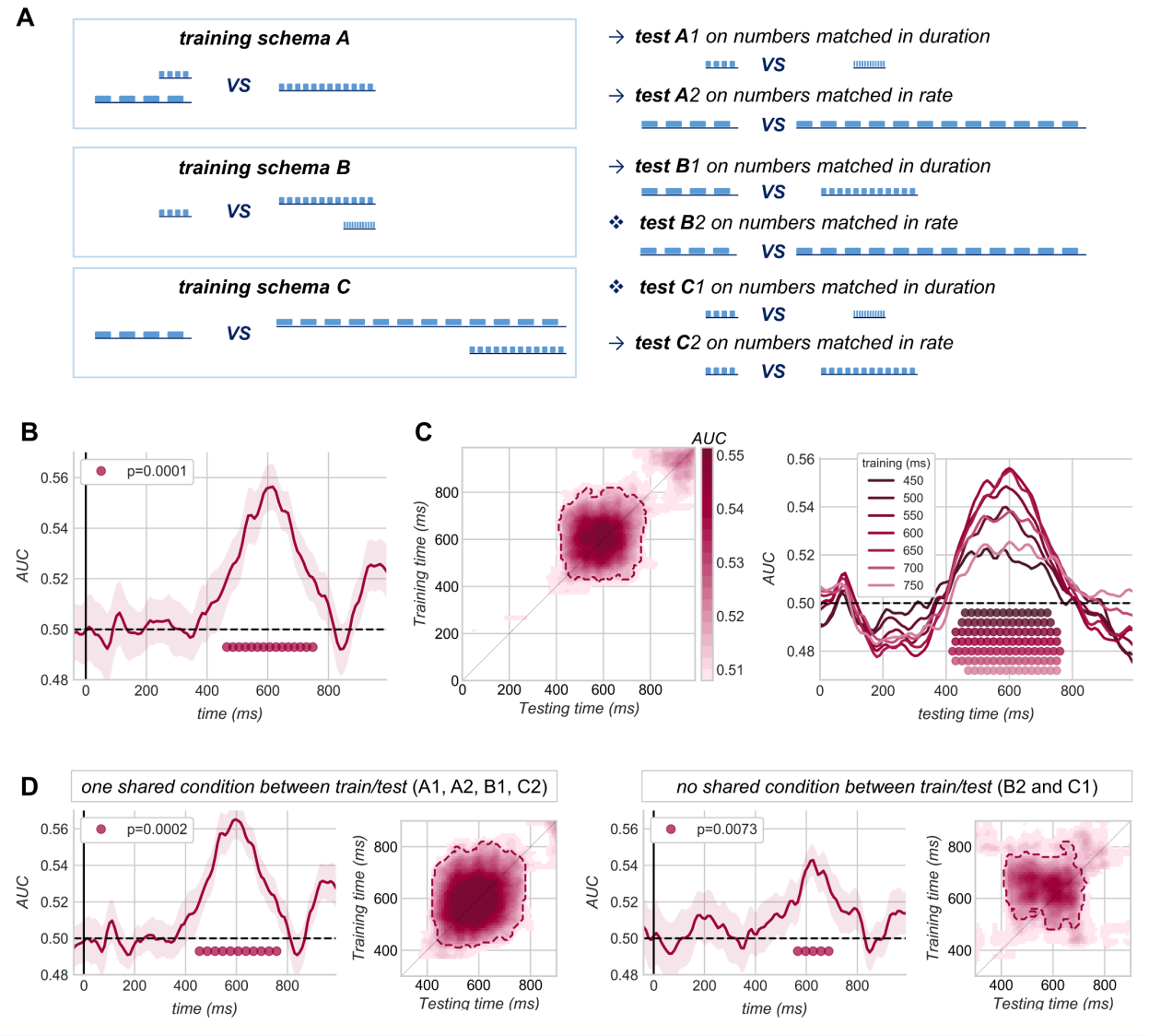

Figure 3.2 Classification of "4" vs "12" from infant neural responses when information on rate and duration is utterly ruled out. (A) Tactical combination of training and test sets. (B) Decoding performance of classifiers trained and tested on 10ms-windows all along the trial: the outcomes of all tests (panel A-right) are averaged. Time 0 corresponds to the onset of the last tone composing the auditory sequences. Shaded areas indicate the standard error (SEM) across subjects (N=26) and dotted black lines mark theoretical chance level. (C) Generalization of the same classifiers across time. Left: the diagonal thin line demarks the scores illustrated in B. Dashed contours delimit statistical significance, calculated by means of a cluster-based permutation t-test against theoretical chance (p=0.0001). The panel on the right offers an alternative visualization of the same generalization scores that highlights how a representative sample of estimators trained in between 440ms and 800ms display overlapping performance dynamics throughout

the trial. (D) Performance dynamics remain similar across tests, even in the most challenging circumstances. Panels on the left depict the averaged outcome of "standard" tests (demarked by arrows in A) performed on pairs of duration-/rate-matched conditions where the non-numerical parameters distinctive of a single class during training were either spread over both classes or removed. Panels on the right report the averaged performance observed for the most critical tests (demarked by diamonds in A), where non-numerical parameters distinctive of a single class during training were assigned only to the alternative class or removed.

To validate our decoding approach, we trained a new set of estimators to separate pairs of conditions distinguished in tone rate but matched for sequence duration ("4L" vs "12M") or differing in total duration but matched for rate ("4M" vs "12M"). We then assessed their performance when the numerical distinctions characterizing the training sets were inverted ("12L" vs "4M" and "12S" vs "4L" respectively), thereby isolating classification scores attributable to rate and duration exclusively. We found no overlap between rate decoding and number classifiability as observed on duration-matched test sets (Figure 3.5A) and no overlap between duration decoding and the performance yielded by number classifiers on rate-matched test sets (Figure 3.5B). These observations confirmed that the successful classification attained in the main analysis (Figure 3.2) is not driven by the fact that, *on average*, sequences of 4 tones are characterized by a slower rate or a shorter total duration.

So far, we have demonstrated that "4" and "12" trials can be reliably discerned from infant neural responses, once the effects related to specific non-numerical parameters are canceled out. Still, correct classifications might be driven by a generalized magnitude/intensity code, where numerical and non-numerical information are integral (Walsh, 2003). Given this eventuality, we used a Representation Similarity Analysis (RSA; Kriegeskorte, 2008) to ask whether (and when) the various quantitative dimensions characterizing the stimuli can be effectively disentangled. At every time point from sequence offset onwards (i.e. over the same window used for decoding), we assessed the correlational distance between the average responses evoked by each pair of auditory conditions. We then used multiple linear regression to model the resulting neural (dis)similarity matrices as a linear combination of three theoretical matrices depicting the (dis)similarity of the sequences along their defining quantitative dimensions: number, rate, duration (Figure 3.3A-top). With this approach, we obtained three series of beta weights reflecting the portion of the neural variance that each quantitative dimension explained *independently* from the other two. Figure 3.3A shows how rate, duration and number were clearly separable: significantly above-zero beta weights imply that, at a certain point within the trial, one quantity modulated neural activity over and beyond the remaining two. Crucially, number exerted the strongest degree of contribution over a relatively late time-window (Figure 3.3A-bottom), with a peak in beta weights observed at 600ms (N=26, 525-805ms: p_{clust}=0.0001). Being fully congruent with the time-course of the classification performance (Figure 3.2B-D), this finding conclusively elucidates that the infant brain estimates numerosity separately from the other magnitudes and in a completely automatic fashion.

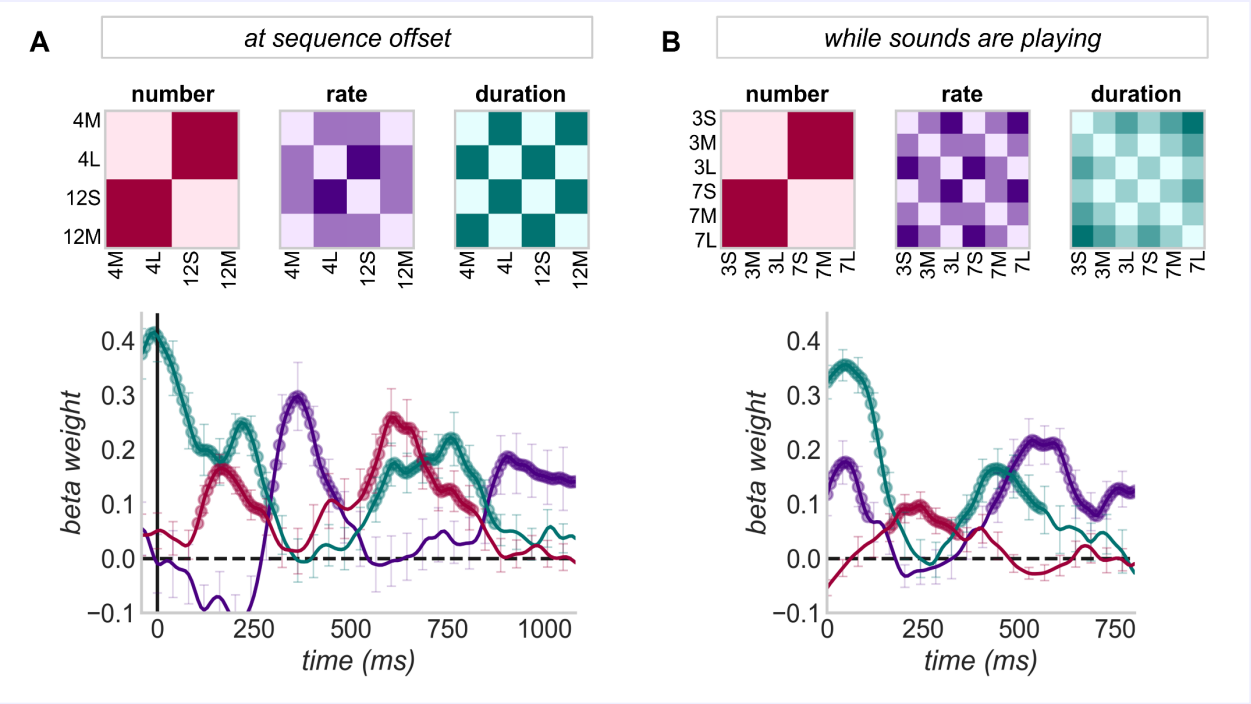

Figure 3.3 RSA uncovers separate encoding of quantitative dimensions and online update. Top: schematic illustration of theoretical matrices depicting the logarithmic distance between pairs of sub-conditions along their defining quantitative dimensions. Darker shades correspond to greater distance/dissimilarity. These three matrices were entered together in a multiple linear regression to explain the corresponding neural distances at each time point. Bottom: standardized beta weights are averaged across subjects (N=26; vertical lines indicate the SEM) and marked by filled circles when significantly above zero (at sequence offset – rate: 295-465ms and from 855ms onwards, p_{clust}= 0.002; duration: -40-295 and 550-865ms, p_{clust}=0.0001; while sounds are playing – rate: 0-105ms and from 425ms onwards, p_{clust}< 0.01; duration: 0-160 and 345-550ms, p_{clust}=0.0001). Beta coefficients for number reveal the presence of two distinct effects: an early modulation is observable from both graphs whereas a later effect, resembling that illustrated in Figure 3.2B, is retrieved only after sounds are finished. This double-peaked pattern could reflect an accumulator mechanism that updates online, which is followed by a final numerical estimation.

Further, number coefficients resulted significantly above zero over an earlier window (N=26, 105-280ms: p_{clust}=0.0001), indicating the existence of a preliminary numerical process not captured by the decoding analysis. Being very close to the final tone of the sequence, this early modulation might reflect a (still) ongoing quantification process corresponding to an accumulator mechanism (Meck & Church, 1983). The latter might consist of a numerosity-sensitive rather than number-selective phase (Verguts & Fias, 2004): at this stage numerical information might be heavily intermingled with the other parameters and thus "discarded" by the strict training-test strategy presented above. To test this interpretation, we investigated the representational similarity structure embedded in the neural signal while the tones composing the sequences were still playing. We reasoned that since an accumulator would update online, the corresponding neural effect should be discernible throughout the stimulation. Conversely, numerosity estimation might occur only after the sequence of tones has

terminated. Restricting the focus on "12" trials, we computed the correlational distances between the average neural signal recorded from the onset of the third and seventh notes (Figure 3.3B-top), then applied the same multiple regression in order to assess/disentangle the modulation of rate, duration (in this case: time elapsed from sequence onset) and number. The choice of "3" vs "7" enabled to minimize multicollinearity between predictors and the risk of low-level confounds (due to e.g. onset and offset proximity) while keeping a reasonable distance between the to-be-compared numerosities (ratio>1:2); thereby providing a set of beta weights that could be conceptually put in relation to the previous. Corroborating our interpretation, beta coefficients for number were significantly positive in between 160 and 335ms (Figure 3.3B-bottom; p_{clust}=0.0007). A similar modulatory effect exerted by the number regressor was observed for "4" vs "9" but not "3" vs "5" (Figure 3.6), revealing that this result is not attributable to the intervention of an object-tracking system (for 3 but not 7; Feigenson et al., 2004; Hyde & Spelke, 2011) and that the accumulator mechanism is imprecise (i.e. a ratio <1:2 is insufficient to discern a numerosity effect).

3.4.1. Testing the abstractness of the infant neural code for numerosity

Our set of results is coherent in demonstrating that the infant brain treated numerosity as a basic property of the auditory sequences, not reducible to other non-numerical variables. This finding is somehow counterintuitive: after all, number encapsulates a *discretization* process, just as rate, and a *cumulative* aspect, just as duration. Speaking of, tone rate and sequence duration did shape neural activity patterns significantly (Figure 3.3). What is the benefit of a primary neural mechanism specifically dedicated to the *accumulation of discretized sensory evidence*? The answer might rely in the unique representational flexibility numerosity affords: unlike the other quantitative parameters, number can be abstracted away from sensory modalities, time and space (Cantlon, 2018). Thus, as a natural continuation of our study, we asked whether the infant brain employs the same numerosity code irrespective of sensory modality, presentation format, and vigilance/attentional state. To this aim, we selected the algorithms that proved successful in isolating a number-specific neural pattern (Figure 3.2B-C) and optimized their learning by training them iteratively on all possible auditory schemas (Figure 3.2A, see *Methods*, section 3.2.5). We then assessed their performance on trials when infants *attended* to *visual* displays of "4" and "12" (in the form of *simultaneous* sets of colorful objects Figure 3.1B). Strikingly, estimators trained in between 440 and 610ms from the last tone composing the auditory sequences performed reliably above chance not only within the auditory modality (Figure 3.2B-C) but also on visual trials (Figure 3.4, N=16: p_{clust}=0.005), with the best performance achieved at 370ms after image onset (e.g. for the classifier trained at 500ms: mean AUC=0.588±0.076). Analogously to what observed on auditory trials, successful classifiers yielded similar performance dynamics, with above-chance scores obtained from ~300 to 580ms relatively to image onset (Figure 3.4B). Such an outcome was not dependent on the type of control used over non-numerical visual attributes, since the scores achieved on trials with fixed extensive parameters (N=16; training 490-510ms, test 360-380ms: mean AUC=0.565±0.092) resulted comparable to those

obtained when intensive parameters were equated across numerosities (mean AUC=0.605±0.148; t=-0.853, p=0.41).

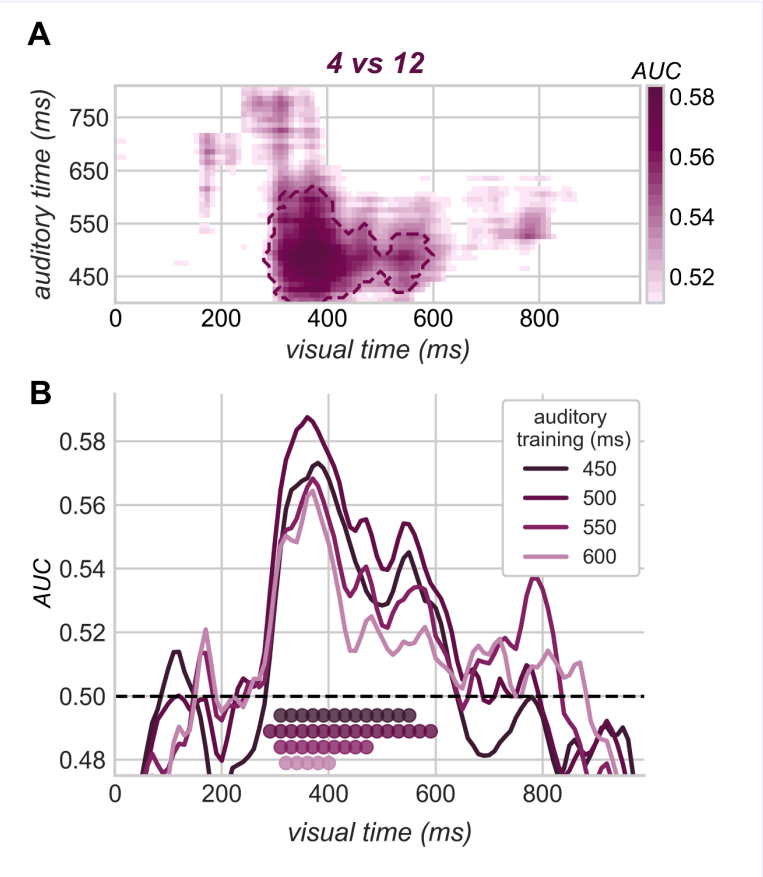

Figure 3.4 Cross-condition decoding reveals an abstract code for number in the preverbal infant brain. (A) Performance of classifiers trained on selected time-windows of auditory trials (y-axis) and tested all along the visual ERPs (x-axis). Dashed contours delimit statistical significance (cluster-based permutation t-test). (B) Alternative visualization of the same generalization scores displayed in A, which illustrates how successful estimators yielded overlapping performance dynamics not only on auditory (Figure 3.2C-right) but also on visual trials.

3.5. DISCUSSION

In the current investigation we exposed preverbal infants to an auditory space composed of sequences of natural sounds embedding a balanced calibration of numerical and non-numerical quantitative parameters. The stimulus space was designed in combination with a strategic multivariate analysis plan which allowed to isolate purely numerical processes from any modulatory effect ascribable to the other quantitative characteristics of the stimuli (single tone duration and inter-tone intervals, rate, sequence duration and total amount of sound/silence). Our pattern of results reveals that the infant brain encodes the numerical information embedded in the auditory sequences separately from all the other dimensions and in full automaticity, i.e. during sleep and within a randomized presentation order.

The fact that the preverbal brain regards discrete quantity as a primary attribute of sound sequences suggests that numerosity is a fundamental, key dimension to represent the external environment. Its importance might derive from the fact that numerical computations allow to merge the disparate

vocabularies of distinct sensory modalities and spatiotemporal axes into a single, informative and flexible descriptor. Coherently with this conception, our investigation revealed that at ~12 weeks of age our brain engages a neural mechanism for numerosity processing that transcends wakefulness state, sense organs and temporal/spatial distribution. This result, impressive and yet barely surprising, recalls two seminal reports pertaining to newborns. Namely, after a brief familiarization phase, 0 to 4 day-olds prefer to look at visual arrays that are matched for number of items with the auditory sequence they hear, provided that the numerical ratio between test displays is at least 1:3 (Izard et al., 2009; Coubart et al., 2014). A tendency of this sort could reflect an instinctive mapping between two arbitrary quantitative dimensions, such as rate and density, similar to those observed in both neonates (de Hevia et al., 2014) and older infants (Lourenco & Longo, 2010; de Hevia & Spelke, 2010). If that was the case, detection of a correspondence between auditory temporally-distributed information and visual ensembles might arise from a generalized magnitude representation (Newcombe et al., 2015). Conversely, the neural evidence provided by the current study corroborates the alternative interpretation according to which newborns can detect a supramodal genuinely numerical correspondence; that is: they "perceive abstract numbers" (Izard et al., 2009). Intriguingly, the observations brought by these behavioral experiments indicate that, although the subjects tested in the present study were 3 months old, the abstract neural code isolated by our analyses is likely to be operational since birth. This possibility is further supported by recent results obtained with hierarchical deep neural networks (DNNs). In the complete absence of learning, tuning to numerosity emerges spontaneously (Zorzi & Testolin, 2018) and enables the networks to succeed in number discrimination tasks even in the presence of incongruent non-numerical quantitative parameters (Kim et al., 2021). This computational phenomenon suggests that the neural code for numerosity isolated here might arise from the interplay between the intrinsic structure of our nervous system and the properties of the external environment. Not only, inspecting the internal dynamics of the aforementioned neural networks disclosed the existence of number-sensitive and number-selective response profiles (Zorzi & Testolin, 2018), whereby tuning to numerosity emerges from monotonic increases and decreases of neuronal activity in the earlier layers of the network (Kim et al., 2021). It is captivating to notice how such "summation coding" is consistent with the accumulator mechanism captured by our RSA analysis (Figure 3.3).

To our knowledge, an abstract neural code for non-symbolic, non-verbal numerosity as that isolated by our classifiers (i.e. a code that transcends format, modality and arousal state) has never been observed in humans before. Two separate fMRI studies on adults have reported overlapping neural activations in response to visually and auditorily presented sequential numerical displays (Piazza et al., 2006) and to sequentially and simultaneously presented visual numerical displays (Dormal et al., 2010). However, subsequent fMRI investigations failed to replicate these findings, suggesting a role for active comparison processes in the previous results (Cavdaroglu et al., 2015; Cavdaroglu & Knops, 2019). Further, the retrieval of a supramodal neural code in infants might appear at odds with the observation that visual numerosity is processed directly, at the level of early cortical areas in adults

(Park et al., 2016; Fornaciai et al., 2017; DeWind et al., 2019; Lucero et al., 2020). As far as current neuroimaging evidence allows to infer, it might possible that initially generalized numerical approximations become modality and format specific as we grow older and more experienced. The plausibility of this hypothesis is supported by the fact that development is characterized by an initial period of hyper-connectivity between sensory brain regions which is subsequently skimmed through phases of retraction and reweighting in the connections (Fransson et al., 2011; Wagner & Dobkins, 2011). Yet, despite the intuitive tendency to conceive low-level cortical mechanisms as putatively sensory-specific, the existence of cross-modal processing at early cortical stages has been documented not only in infants (Werchan et al., 2018) but also (and mainly) within adulthood (Murray et al., 2016). For instance, multisensory convergence has been observed with adult fMRI: auditory stimulation alone can trigger robust responses in primary visual cortices and simple checkerboards activate primary auditory areas (Martuzzi et al., 2007). Using EEG on adults, contralateral ERPs over the occipital scalp (with sources located in extrastriate regions) have been retrieved within purely auditory experimental stimulation when sounds were not relevant to the task (McDonald et al., 2013). Concerning adult structural connectivity, there are fiber tracts forming a direct pathway between the Heschl's gyrus and the occipital pole as well as anterior portions of the calcarine sulcus (Beer et al., 2011). Beyond the consideration that supramodal processes can occur at all cortical stages throughout life (Ghazanfar & Schroeder, 2006; Foxe & Schroeder, 2005), the most meaningful insights to the aim of the current discussion derive from psychophysics. As mentioned in the introduction of this chapter, numerosity estimates in adults are susceptible to adaptation (Burr & Ross, 2008). Impressively, adaptation to numerosity generalizes across formats and sensory modalities: from sequential streams to spatial arrays (and vice versa), from auditory sequences to visual displays (always in a bidirectional fashion). Crucially, cross-format and cross-modal adaptation effects are nearly as large as those observed within-format and within-modality, revealing that to be adapting is an abstract quantity system (Arrighi et al., 2014). Further, adaptation remains spatially specific in all cases, suggesting that to be involved is a relatively basic encoding mechanism rather than a higher-level cognitive construct (Burr et al., 2018). Overall, this pattern of behavioral findings provides a glimpse of a discrete quantitative system in adults that parallels in all ways that isolated here, within a developmental neuroimaging setting. Thus, our observations are ultimately coherent with and further extend the evidence from adults, showing the existence, early in life, of a spontaneous, specialized and modality-independent neural mechanism that encodes numerosity in the form of a basic sensory descriptor; that is to say: the existence of a *primitive and abstract number sense*.

3.6. SUPPLEMENTARY MATERIALS

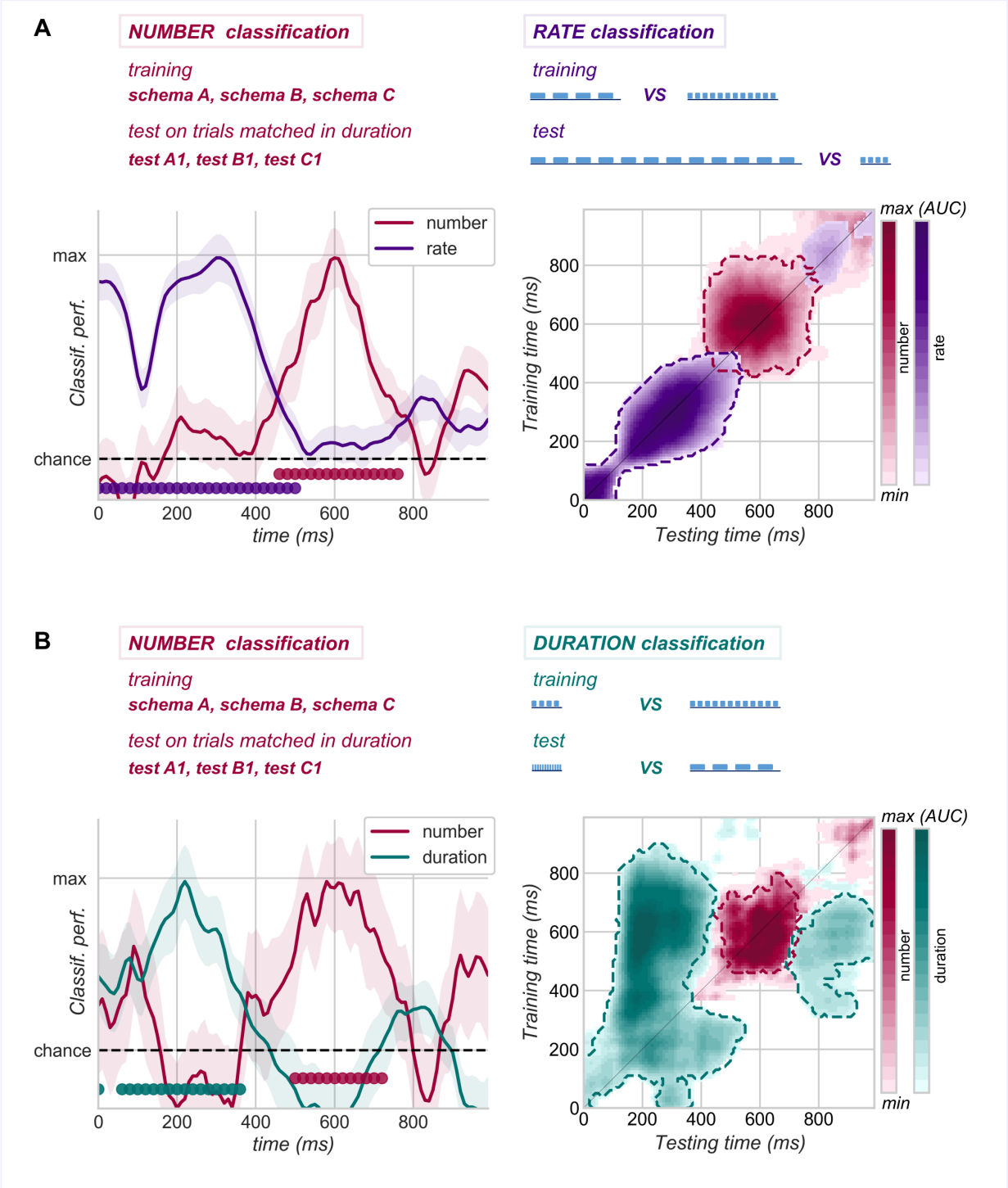

Figure 3.5 Sanity-check analyses (complement of Figure 3.2) (A) Within the training sets employed in our main decoding analysis, class "4" was necessarily characterized by auditory sequences that are on average slower relatively to those corresponding to class "12". With this qualitative assessment, we asked whether above-chance performance on conditions matched for duration (i.e. when duration-related effects

are completely uninformative, see Figure 3.2A for an intuitive illustration) might reflect the retrieval of a "slower than/faster than" type of computation. In other words, we asked whether successful classification of numerosity could actually derive from rate distinctions. In the two panels, rate-based classification performance is superimposed over the average scores obtained within the main decoding analysis precisely when classifiers were tested on pairs of conditions that differed for both rate and numerosity. We found no overlap between the two performances, indicating that the main classification scores were not contaminated by the retrieval of rate-related effects. (B) Within our main training sets, class "4" was necessarily characterized by auditory sequences that are on average shorter relatively to those corresponding to class "12". This qualitative assessment mirrors that presented in A. Namely, we tested whether successful performance on conditions matched for rate (i.e. when rate-related effects are completely uninformative) could reflect the fact that the infant brain encoded the stimuli in terms of "shorter than/longer than". As before, the superimposition of classification scores reflecting pure duration-related effects on the performance obtained in those main tests that included between-class duration discrepancies clearly testifies how the scores observed in the latter were not contaminated by duration-related processing.

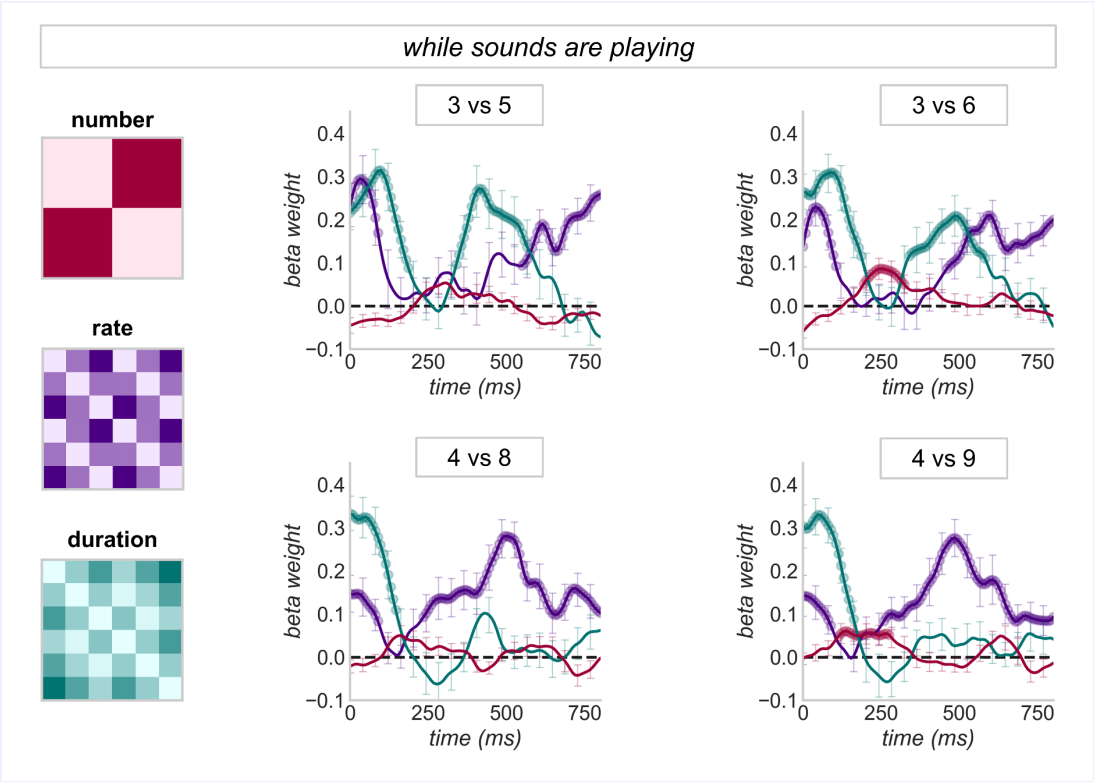

Figure 3.6 RSA within sequence indicates that online numerical accumulation is imprecise. Multiple regression analyses mirroring that illustrated in Figure 3.3B in all aspects except for the numerical contrast under investigation. Standardized beta weights are averaged across subjects (N=26; vertical lines indicate the SEM) and marked by filled circles when significantly above zero (3vs5 – rate: 0-90ms pclust=0.0001 and from 545ms onwards p_{clust}= 0.01; duration: 0-185 and 335-560ms, p_{clust}=0.0001 | 3vs6 – number: 190-320ms p_{clust}=0.0054; rate: 0-105ms and from 455ms onwards, p_{clust}< 0.01; duration: 0-185 and 335-575ms, p_{clust}=0.0001 | 4vs8 – rate: 0-90ms p_{clust}=0.0001 and from 230ms onwards p_{clust}= 0.0035; duration: 0-145 p_{clust}=0.0005 | 4vs9 – number: 110-270ms p_{clust}=0.007; rate: 0-90ms and from 256ms onwards, p_{clust}< 0.01; duration: 0-160 p_{clust}=0.0001).

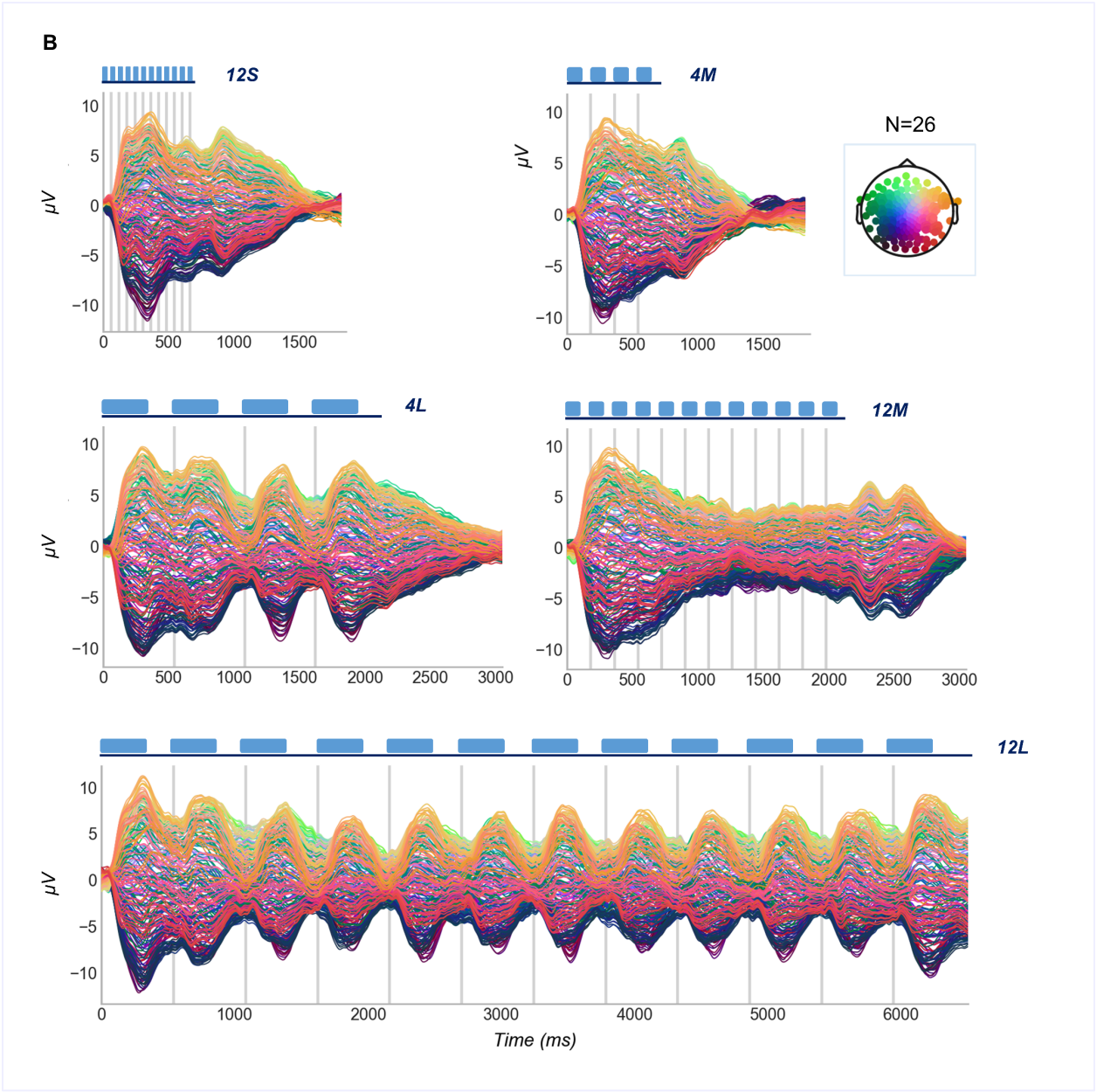

Figure 3.7 Grand average ERPs to auditory sequences

3.7. References

Arrighi, R., Togoli, I., & Burr, D. C. (2014). A generalized sense of number. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1797), 20141791. https://doi.org/10.1098/rspb.2014.1791

Aulet, L. S., & Lourenco, S. F. (2021a). The relative salience of numerical and non-numerical dimensions shifts over development: A re-analysis of Tomlinson, DeWind, and Brannon (2020). *Cognition*, *210*, 104610. https://doi.org/10.1016/j.cognition.2021.104610

Aulet, L. S., & Lourenco, S. F. (2021b). *No intrinsic number bias: Evaluating the role of perceptual discriminability in magnitude categorization* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/eh5pb

Beer, A. L., Plank, T., & Greenlee, M. W. (2011). Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex. *Experimental Brain Research*, *213*(2), 299. https://doi.org/10.1007/s00221-011-2715-y

Borghesani, V., de Hevia, M. D., Viarouge, A., Pinheiro-Chagas, P., Eger, E., & Piazza, M. (2019). Processing number and length in the parietal cortex: Sharing resources, not a common code. *Cortex*, *114*, 17–27. https://doi.org/10.1016/j.cortex.2018.07.017

Brannon, E. M., Lutz, D., & Cordes, S. (2006). The development of area discrimination and its implications for number representation in infancy. *Developmental Science*, *9*(6), F59–F64. https://doi.org/10.1111/j.1467-7687.2006.00530.x

Brannon, E. M., Suanda, S., & Libertus, K. (2007). Temporal discrimination increases in precision over development and parallels the development of numerosity discrimination. *Developmental Science*, *10*(6), 770–777. https://doi.org/10.1111/j.1467-7687.2007.00635.x

Burr, D. C., Anobile, G., & Arrighi, R. (2018). Psychophysical evidence for the number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170045. https://doi.org/10.1098/rstb.2017.0045

Burr, D. C., & Ross, J. (2008). A Visual Sense of Number. *Current Biology*, *18*(6), 425–428. https://doi.org/10.1016/j.cub.2008.02.052

Butterworth, B. (2018). The implications for education of an innate numerosity-processing mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170118. https://doi.org/10.1098/rstb.2017.0118

Cantlon, J. F. (2018). How Evolution Constrains Human Numerical Concepts. *Child Development Perspectives*, *12*(1), 65–71. https://doi.org/10.1111/cdep.12264

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in Cognitive Sciences*, *13*(2), 83–91. https://doi.org/10.1016/j.tics.2008.11.007

Cantrell, L., & Smith, L. B. (2013). Open questions and a proposal: A critical review of the evidence on infant numerical abilities. *Cognition*, *128*(3), 331–352. https://doi.org/10.1016/j.cognition.2013.04.008

Castaldi, E., Piazza, M., Dehaene, S., Vignaud, A., & Eger, E. (2019). Attentional amplification of neural codes for number independent of other quantities along the dorsal visual stream. *ELife*, *8*, e45160. https://doi.org/10.7554/eLife.45160

Cavdaroglu, S., Katz, C., & Knops, A. (2015). Dissociating estimation from comparison and response eliminates parietal involvement in sequential numerosity perception. *NeuroImage*, *116*, 135–148. https://doi.org/10.1016/j.neuroimage.2015.04.019

Cavdaroglu, S., & Knops, A. (2019). Evidence for a Posterior Parietal Cortex Contribution to Spatial but not Temporal Numerosity Perception. *Cerebral Cortex*, *29*(7), 2965–2977. https://doi.org/10.1093/cercor/bhy163

Coubart, A., Izard, V., Spelke, E. S., Marie, J., & Streri, A. (2014). Dissociation between small and large numerosities in newborn infants. *Developmental Science*, *17*(1), 11–22. https://doi.org/10.1111/desc.12108

de Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., & Streri, A. (2014). Representations of space, time, and number in neonates. *Proceedings of the National Academy of Sciences*, *111*(13), 4809–4813. https://doi.org/10.1073/pnas.1323628111

de Hevia, M. D., & Spelke, E. S. (2010). Number-Space Mapping in Human Infants. *Psychological Science*, *21*(5), 653–660. https://doi.org/10.1177/0956797610366091

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, USA.

DeWind, N. K., Park, J., Woldorff, M. G., & Brannon, E. M. (2019). Numerical encoding in early visual cortex. *Cortex*, *114*, 76–89. https://doi.org/10.1016/j.cortex.2018.03.027

Dormal, V., Andres, M., Dormal, G., & Pesenti, M. (2010). Mode-dependent and mode-independent representations of numerosity in the right intraparietal sulcus. *NeuroImage*, *52*(4), 1677–1686. https://doi.org/10.1016/j.neuroimage.2010.04.254

Edwards, L. A., Wagner, J. B., Simon, C. E., & Hyde, D. C. (2016). Functional brain organization for number processing in pre-verbal infants. *Developmental Science*, *19*(5), 757–769. https://doi.org/10.1111/desc.12333

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*(Aug), 1871–1874.

Feigenson, L. (2007). The equality of quantity. *Trends in Cognitive Sciences*, *11*(5), 185–187. https://doi.org/10.1016/j.tics.2007.01.006

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Fornaciai, M., Brannon, E. M., Woldorff, M. G., & Park, J. (2017). Numerosity processing in early visual cortex. *NeuroImage*, *157*, 429–438. https://doi.org/10.1016/j.neuroimage.2017.05.069

Foxe, J. J., & Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing: *NeuroReport*, *16*(5), 419–423. https://doi.org/10.1097/00001756-200504040-00001

Fransson, P., Åden, U., Blennow, M., & Lagercrantz, H. (2011). The Functional Architecture of the Infant Brain as Revealed by Resting-State fMRI. *Cerebral Cortex*, *21*(1), 145–154. https://doi.org/10.1093/cercor/bhq071

Georges, C., Guillaume, M., & Schiltz, C. (2020). A robust electrophysiological marker of spontaneous numerical discrimination. *Scientific Reports*, *10*(1), 18376. https://doi.org/10.1038/s41598-020-75307-y

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–285. https://doi.org/10.1016/j.tics.2006.04.008

Gordon, P. (2004). Numerical Cognition Without Words: Evidence from Amazonia. *Science*, *306*(5695), 496–499. https://doi.org/10.1126/science.1094492

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

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

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2016). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Hamamouche, K., & Cordes, S. (2019). Number, time, and space are not singularly represented: Evidence against a common magnitude system beyond early childhood. *Psychonomic Bulletin & Review*, *26*(3), 833–854. https://doi.org/10.3758/s13423-018-1561-3

Harvey, B. M., Fracasso, A., Petridou, N., & Dumoulin, S. O. (2015). Topographic representations of object size and relationships with numerosity reveal generalized quantity processing in human parietal cortex. *Proceedings of the National Academy of Sciences*, *112*(44), 13525–13530. https://doi.org/10.1073/pnas.1515414112

Hyde, D. C., Boas, D. A., Blair, C., & Carey, S. (2010). Near-infrared spectroscopy shows right parietal specialization for number in pre-verbal infants. *NeuroImage*, *53*(2), 647–652. https://doi.org/10.1016/j.neuroimage.2010.06.030

Hyde, D. C., & Spelke, E. S. (2011). Neural signatures of number processing in human infants: Evidence for two core systems underlying numerical cognition. *Developmental Science*, *14*(2), 360–371. https://doi.org/10.1111/j.1467-7687.2010.00987.x

Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct Cerebral Pathways for Object Identity and Number in Human Infants. *PLOS Biology*, *6*(2), e11. https://doi.org/10.1371/journal.pbio.0060011

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, *106*(25), 10382–10385. https://doi.org/10.1073/pnas.0812142106

Kim, G., Jang, J., Baek, S., Song, M., & Paik, S.-B. (2021). Visual number sense in untrained deep neural networks. *Science Advances*, *7*(1), eabd6127. https://doi.org/10.1126/sciadv.abd6127

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

King, J.-R., Pescetelli, N., & Dehaene, S. (2016). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. *Neuron*, *92*(5), 1122–1134. https://doi.org/10.1016/j.neuron.2016.10.051

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Leibovich, T., & Henik, A. (2013). Magnitude processing in non-symbolic stimuli. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00375

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16000960

Libertus, M. E., & Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Developmental Science*, *13*(6), 900–906. https://doi.org/10.1111/j.1467-7687.2009.00948.x

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of Large and Small Numerosities by Human Infants. *Infancy*, *5*(3), 271–290. https://doi.org/10.1207/s15327078in0503_2

Lourenco, S. F., & Longo, M. R. (2010). General Magnitude Representation in Human Infants. *Psychological Science*, *21*(6), 873–881. https://doi.org/10.1177/0956797610370158

Lucero, C., Brookshire, G., Sava-Segal, C., Bottini, R., Goldin-Meadow, S., Vogel, E. K., & Casasanto, D. (2020). Unconscious Number Discrimination in the Human Visual System. *Cerebral Cortex*, *30*(11), 5821–5829. https://doi.org/10.1093/cercor/bhaa155

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

Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J.-P., Maeder, P. P., Clarke, S., & Meuli, R. A. (2007). Multisensory Interactions within Human Primary Cortices Revealed by BOLD Dynamics. *Cerebral Cortex*, *17*(7), 1672–1679. https://doi.org/10.1093/cercor/bhl077

McDonald, J. J., Störmer, V. S., Martinez, A., Feng, W., & Hillyard, S. A. (2013). Salient Sounds Activate Human Visual Cortex Automatically. *Journal of Neuroscience*, *33*(21), 9194–9201. https://doi.org/10.1523/JNEUROSCI.5902-12.2013

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 320–334. https://doi.org/10.1037/0097-7403.9.3.320

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). Multiple cues for quantification in infancy: Is number one of them? *Psychological Bulletin*, *128*(2), 278–294. https://doi.org/10.1037/0033-2909.128.2.278

Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., & Matusz, P. J. (2016). The multisensory function of the human primary visual cortex. *Neuropsychologia*, *83*, 161–169. https://doi.org/10.1016/j.neuropsychologia.2015.08.011

Newcombe, N. S., Levine, S. C., & Mix, K. S. (2015). Thinking about quantity: The intertwined development of spatial and numerical cognition. *WIREs Cognitive Science*, *6*(6), 491–505. https://doi.org/10.1002/wcs.1369

Park, J. (2018). A neural basis for the visual sense of number and its development: A steady-state visual evoked potential study in children and adults. *Developmental Cognitive Neuroscience*, *30*, 333–343. https://doi.org/10.1016/j.dcn.2017.02.011

Park, J., DeWind, N. K., Woldorff, M. G., & Brannon, E. M. (2016). Rapid and Direct Encoding of Numerosity in the Visual Stream. *Cerebral Cortex*, *26*(2), 748–763. https://doi.org/10.1093/cercor/bhv017

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

Piantadosi, S. T., & Cantlon, J. F. (2017). True Numerical Cognition in the Wild. *Psychological Science*, *28*(4), 462–469. https://doi.org/10.1177/0956797616686862

Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences*, *14*(12), 542–551. https://doi.org/10.1016/j.tics.2010.09.008

Piazza, M., Mechelli, A., Price, C. J., & Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: An fMRI study. *Brain Research*, *1106*(1), 177–188. https://doi.org/10.1016/j.brainres.2006.05.104

Pica, P. (2004). Exact and Approximate Arithmetic in an Amazonian Indigene Group. *Science*, *306*(5695), 499–503. https://doi.org/10.1126/science.1102085

Rousselle, L., Palmers, E., & Noël, M.-P. (2004). Magnitude comparison in preschoolers: What counts? Influence of perceptual variables. *Journal of Experimental Child Psychology*, *87*(1), 57–84. https://doi.org/10.1016/j.jecp.2003.10.005

Smyth, R. E., & Ansari, D. (2020). Do infants have a sense of numerosity? A p-curve analysis of infant numerosity discrimination studies. *Developmental Science*, *23*(2), e12897. https://doi.org/10.1111/desc.12897

Soltész, F., Szűcs, D., & Szűcs, L. (2010). Relationships between magnitude representation, counting and memory in 4- to 7-year-old children: A developmental study. *Behavioral and Brain Functions*, *6*(1), 13. https://doi.org/10.1186/1744-9081-6-13

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Van Rinsveld, A., Guillaume, M., Kohler, P. J., Schiltz, C., Gevers, W., & Content, A. (2020). The neural signature of numerosity by separating numerical and continuous magnitude extraction in visual cortex with frequency-tagged EEG. *Proceedings of the National Academy of Sciences*, *117*(11), 5726–5732. https://doi.org/10.1073/pnas.1917849117

VanMarle, K., & Wynn, K. (2006). Six-month-old infants use analog magnitudes to represent duration. *Developmental Science*, *9*(5), F41–F49. https://doi.org/10.1111/j.1467-7687.2006.00508.x

Verguts, T., & Fias, W. (2004). Representation of Number in Animals and Humans: A Neural Model. *Journal of Cognitive Neuroscience*, *16*(9), 1493–1504. https://doi.org/10.1162/0898929042568497

Wagner, K., & Dobkins, K. R. (2011). Synaesthetic Associations Decrease During Infancy. *Psychological Science*, *22*(8), 1067–1072. https://doi.org/10.1177/0956797611416250

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Werchan, D. M., Baumgartner, H. A., Lewkowicz, D. J., & Amso, D. (2018). The origins of cortical multisensory dynamics: Evidence from human infants. *Developmental Cognitive Neuroscience*, *34*, 75–81. https://doi.org/10.1016/j.dcn.2018.07.002

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11. https://doi.org/10.1016/S0010-0277(99)00066-9

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88–101. https://doi.org/10.1111/j.1467-7687.2005.00395.x

Zorzi, M., & Testolin, A. (2018). An emergentist perspective on the origin of number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170043. https://doi.org/10.1098/rstb.2017.0043

Chapter 4. THE NEURAL REALITY OF PITCH CHROMA IN EARLY INFANCY

ABSTRACT

At the physical level the experience of pitch has a single determinant: the repetition rate of a waveform in the acoustic signal. Yet, psychologists have described pitch as composed of two perceptual dimensions, height and chroma. According to opponents of an ongoing controversy, chroma might correspond to either a basic perceptual property dependent on biological constraints or a higher-order cognitive construct crafted by culture. Here we used high-density Electroencephalography (EEG) and multivariate-pattern analyses to characterize pitch processing in humans at 3 months of age. We found that, when exposed to repetitive sequences of orchestral tones, infants encode both pitch height and pitch chroma in a completely automatic fashion and with neatly divergent dynamics. Our classifiers were able to isolate height-specific information from the neural signal right after the onset of the auditory sequences; beyond ~600ms the performance of height decoders fell at chance and never recovered. On the other hand, neural patterns specific for chroma could be retrieved later in the trial, over multiple time windows throughout the unfolding of the auditory sequence and after sequence offset. Overall, this study demonstrates that not only pitch height but also pitch chroma constitute a basic organizing principle of neural responsivity very early in development. We speculate that separate encoding mechanisms reflect distinct functional roles carried by the two dimensions.

4.1. INTRODUCTION

Pitch is one of the fundamental aspects of sound as it carries information upon source identity, prosody and melody. To convey meaning, both speech and music rely on pitch patterns and pitch relations (e.g. Curtis & Bharucha, 2010). Developmentally, learning to link the latter to different messages is likely to be one of the first steps for infants to make sense of sounds (Fernald, 1989).
The experience of pitch is related to the frequency of vibrations, in the air, hitting the eardrum. Sensitivity to frequency is extremely precocious. For instance, at 36-39 weeks of gestation, fetuses repetitively exposed to a given piano tone, react with a heart rate deceleration when the constituent frequencies of the tone are changed (Lecanuet et al., 2000). At ~35 weeks of gestation, occasional changes in the frequency components of synthetic sounds elicit mismatch neural responses traceable with MEG (Draganova et al., 2005). These observations indicate that pitch-related information can be differentiated already in utero and further suggest an early cortical involvement in this ability.

Natural sounds eliciting a pitch sensation are usually complex, in that the present energy at a fundamental frequency (F0) and at various harmonics, integer multiples of the latter. Such components are processed separately by the auditory system, within frequency-specific channels. Being unitary, our sensation of pitch is not given by the stimulus itself but is the result of an integrative mechanism implemented relatively late in the processing hierarchy, in regions adjacent to the (adult) primary auditory cortex (Bendor & Wang, 2005). A classical phenomenon that enables to isolate pitch extraction from frequency processing is that of the missing fundamental: the pitch of a complex remains the same in our perception, whether or not the fundamental frequency is actually present within the stimulus. Building on this phenomenon, He & Trainor (2009) have shown that authentic pitch perception might appear only around the 4th month of age: in their EEG study, 3-month olds could not detect changes in the pitch of sounds lacking the fundamental, suggesting that early reactions to sounds are based merely on frequency. Although inconsistent with some behavioral observations (Lau & Werner, 2012), these conclusions are compatible with the documented immaturity of the auditory cortex in the very first months of life (Werner et al., 2012). Yet, it remains highly underspecified how complex sounds are processed by an immature brain: young infants might encode F0, the centroid frequency or their spectral shapes, as all these three factors have been shown to exert separable effects on adult neural responses (Crottaz-Herbette & Ragot, 2000; Cansino et al., 2003; Warren et al., 2005).

At the physical level, pitch has a single determinant: the repetition rate of a sound wave. Nevertheless, for nearly two centuries, psychologists and theoreticians have been depicting pitch as a bi-dimensional attribute. First, there is *height*: a monotonic dimension that goes from low to high as the fundamental frequency increases. The second component, *chroma*, is a cyclical qualitative dimension that repeats each time the fundamental frequency doubles (for example: the fundamental frequencies of 220 and 440Hz both corresponds to the A note in Western music notation). Overall, the perceptual organization of pitch can be modeled as a helix formed by a linear component, height,

and a sinusoidal component, chroma (Shepard, 1982). Such characterization takes into account the so-called "octave equivalence" phenomenon (Figure 4.1): two tones whose fundamental frequencies stand at a 1:2 ratio (i.e. forming an octave interval) are experienced as very similar one another and treated as equivalent by nearly all musical scales aver evolved (Burns, 1999; Apel, 2003).

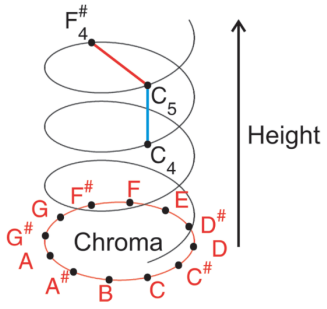

Figure 4.1 Pitch helix. The distance between each 2 points reflects the perceptual similarity between the corresponding notes (Western notation): two sounds separated by an octave (blue line) are closer than two sounds separated by only half-octave (red line) because they share the same chroma. Adapted from Briley et al., 2013.

A few neuroimaging studies, carried on human adults, suggest that the cortical representation of pitch is consistent with the helix model. To start with, the first fMRI investigation on this topic (Warren et al., 2003) has shown that the two dimensions engage the secondary auditory cortex in a differential manner. That is, when orthogonal, changes in pitch height trigger specific activations posterior to Heschl's gyrus (planum temporale) whereas changes in chroma are reflected by specific activity in more anterior regions (planum polare). With EEG, Briley and collaborators (2013) observed that responses to complex tonal sounds adapt in a non-monotonic fashion: in their assessment, neural adaptation was stronger when the adapter and the probe were separated by octave intervals rather than half-octaves. Since the amount of adaptation triggered by an adapter on the probe depends on the overlap between/selectivity of their neural representations, such an effect stands in perfect accordance with the predictions of the helix model.

Yet, more recent evidence calls into question the observations just described. In the EEG investigation conducted by Regev and colleagues (2019) occasional changes of pitch height elicited a mismatch response, whereas changes in pitch chroma did not, despite the latter being promptly detected within a separate behavioral task. According to the authors, this pattern of results indicates that chroma, unlike height, requires high order cognitive functions, such as attention, working memory and learning, in order to be processed. Considering that both the adaptation paradigm employed by Briley et al. (2013) and the MMN protocol adopted by Regev et al. (2019) are supposed to tackle a pre-attentive and automatic encoding of the stimuli, the conclusions from the latter study are at odds with the previous. Conversely, they are compatible with the proposition of several scholars sustaining that octave equivalence is the result of cognitive and cultural factors, rather than a basic perceptual property linked to physiological constraints. In conformity with this idea, both 4- to 9-year-old children and members of the Tsimane tribe (an Amazonia population that lives in isolation from the Western culture) appear insensitive to octave equivalence, indicating that chroma might correspond to a acquired "concept" rather than a fundamental percept (Sergeant, 1983; Jacoby et al., 2019).

Studying sound perception early in infancy has the potential to bring precious clues about this issue. Within the developmental neuroimaging field, a wealth of experiments has investigated mismatch responses to pitch changes starting from the very first days of life (see He et al., 2007 for detailed review). However, the dimensions of height and chroma have never been separated before: in all cases, they were confounded one with the other since the paradigms included stimuli that differed for both.

The purpose of this study is to gain further insights upon the role chroma holds for the human brain, while providing an in-depth characterization of early pitch processing. To address our objectives, we applied multivariate analysis techniques to the neural responses collected from twenty-six 3-month-olds subjects while exposed to sequences of naturally-rich orchestral tones. From the original experiment, designed to query numerosity processing (Figure 3.1A), we selected four types of trials: tones with either a medium (120ms; "M") or a long (360ms; "L") duration, repeated in sequences of either 4 or 12. Sounds corresponded to the notes C3, G3, C4 and G4 and they were played by two string instruments, a cello and a viola. The inclusion of two different timbres within the paradigm was crucial for our goals. In case each note was produced by the same instrument throughout the experiment, there would have been a one-to-one correspondence between fundamental frequencies and spectral profiles, preventing the possibility to disentangle which features the infant brain favors or disregards when encoding the stimuli. Further, when two tones coming from the same source have the same chroma, the harmonics and the fundamental frequency of the higher note are all contained in the spectrum of the lower sound. At the neural level, these sounds might be represented similarly simply due to their shared frequency components. Crucially, different instruments playing the same note give rise to tones with the same fundamental frequency but distinct spectral content. More precisely for our experiment, resonator chambers of different size implied a shift of the energy distribution up to higher frequencies when the viola was playing. Thus, by varying the timbre, we created the opportunity to isolate the fundamental frequency from other spectral characteristics and to query authentic chroma processing by ruling out possible effects arising from mere harmonic overlap.

4.2. MATERIALS & METHODS

4.2.1. Stimuli

The sounds composing the auditory sequences consisted of two musical notes, C and G, sampled from two octaves, the 3rd and the 4th, and played by two string instruments, a viola and a cello, resulting in a total of 8 sub-conditions (2 chromas x 2 octaves x 2 timbres).

The four notes composing the auditory space (C3, G3, C4, G4) were appositely chosen to form perfect intervals (perfect fifth, perfect fourth, octave and perfect twelfth), in light of the perceptual advantage shown for the latter in 6 and 9 mo-old infants (Schellenberg & Trehub, 1996) and taking into account that sensitivity to consonance appears extremely early in life (Trainor et al., 2002). Importantly, the

fact that any possible pair of stimuli formed a perfect interval prevented us from mistaking consonance-related effects for chroma-related ones.

Stimuli were synthesized with Ableton Live 10 (Berlin, Germany), relying on a database of orchestral sounds recorded from musicians of the Boston Ballet and Boston Symphony Orchestras in optimal environmental conditions. Previous studies have shown that the employment of artificial sounds within the experimental paradigm strongly undermines the ecological validity of the results (Hoeschele et al., 2015; Bitterman et al., 2008). Given that our goal was to capture how the infant brain encodes pitch in real-life scenarios, using naturally-rich stimuli was thus essential. At the same time, the use of a digital audio workstation (DAW) enabled to control for low-level acoustical variations that could have potentially compromised the interpretability of our results. For example, all tones were characterized by similar waveforms composed of an attack, stationary portion and decay period with matched shapes and lengths. This characteristic of the auditory space, granted by the use of the DAW, prevented irrelevant features of the envelope from creating spurious differences between sub-conditions.
Tones were synthesized in *pizzicato* with a velocity of 112, corresponding to the dynamic *fortissimo*. Once crafted in Ableton, they were edited with the aid of PRAAT (Boersma & Weenink, 2017) and Audacity (https://www.audacityteam.org). Their precise durations were set to either 360 or 120ms (to obtain a total of 16 unique stimuli: 8 sub-conditions x 2 durations) and their intensity (i.e. root-mean-squared) was scaled at 75dB. Sound offsets were ramped down with a 5ms linear slope to avoid abrupt clicks. The average autocorrelation value of the final stimuli, corresponding to a measure of their pitch strength, was 0.98±0.01. Their spectral characteristics are reported in Table 4.1.

Participants, procedure and EEG data preprocessing are described in sections 3.2.1 and 3.2.3-4

4.2.2. Epochs

EEG data analysis concerned three types of epochs, segmented at different moments in respect to the auditory stimulation. The first analyses were conducted on "onset epochs" spanning from the onset of the first tone composing the sequences until 1080ms. This portion of the trials included either: four medium tones (120ms-long sounds + 60ms inter-tone-interval) and a 360ms silent gap; six medium tones; or two long tones (360ms-long sounds + 180ms inter-tone-interval). To investigate the effect of repetitive stimulation on pitch processing, we used a balanced group of epochs spanning from the onset of the sequence up to 2s and including either eleven medium tones (50% of cases) or four long tones. Lastly, we used "offset epochs" which (in the main version of the analysis) covered the onset of the last tone composing the sequence until 1100ms thereafter.
Before submitting them to the main analyses, all epochs were low-pass filtered at 20Hz and mathematically re-referenced to the mean of all channels.

note	*instrument*	*F0 (Hz)*	*CG (Hz)*	*SD (Hz)*
C3	**cello**	130.8	374.7	106.0
	viola	130.8	347.4	191.1
G3	**cello**	196.0	379.2	117.2
	viola	195.9	421.0	216.9
C4	**cello**	261.7	296.6	166.9
	viola	261.7	389.4	232.4
G4	**cello**	392.0	405.4	107.9
	viola	392.0	554.4	296.9

Table 4.1: spectral parameters defining the experimental sub-conditions.
Each value corresponds to the average of the measurements calculated with Praat (www.praat.org) from 2 stimuli: medium and long notes. The fundamental frequency (F0) is computed through a popular autocorrelation method (Boersma, 1993) that takes the strongest periodic component of several time windows across the stimulus and averages them to yield a single value. The center of gravity (CG) is the spectral centroid, calculated by weighting the mean frequency value of the spectrum by the distribution of signal amplitudes across the spectrum. Higher values coincide with brighter sounds. The standard deviation (SD) is (the square root of) the second central moment of the spectrum and is a measure of spread, i.e. how the spectrum is distributed around its centroid.
The two orchestral strings included in our paradigm are very similar in construction but differ for their size. When playing the same note, these instruments produce sounds characterized by a spectral envelope with the same shape but different scales: a smaller resonator results in higher formant frequencies and wider bandwidths i.e. the energy distribution moves upward along the frequency axis and is more dilated (Dinther & Patterson, 2006). Such a phenomenon is clearly discernible from the CG values within the 4th octave and from the estimations of spectral spread (i.e. SD is systematically lower for tones played by the cello, indicating that their spectrum is more tightly concentrated around the centroid).

4.2.3. Decoding

Multivariate pattern analyses were conducted within subject, relying on the Scikit-Learn (Pedregosa et al., 2011) and MNE (Gramfort et al., 2013, 2014) Python packages. To decode *in time* epochs were always divided into consecutive windows of 10ms, each corresponding to a matrix with the shape *n* channels *x* 5 samples (sampling rate = 500Hz, 5 samples=10ms). Each analysis was carried on a single window with the general aim of predicting a vector of categorical data (*y*) from a matrix of single-trial neural data (*X)* which included all EEG channels. To decode pitch height we used multi-class problems with four labels, corresponding to the identities of the notes forming the auditory space: "C3" vs "G3" vs "C4" vs "G4", whereas chroma decoding entailed binary problems contrasting the classes "C" and "G".

To avoid overfitting, all analyses were performed within a stratified cross-validation procedure composed of 100 loops. At each run, trials were first shuffled and then assigned to training and test sets. For height classification, the size of the two sets was calibrated to guarantee a minimum of two tests on each note and, at the same time, maximize the amount of data provided for learning. The

partitioning was always performed in a stratified fashion such that all sources of both relevant and irrelevant variability (e.g. notes but also single tone duration) were distributed in equal proportions. In cross-instrument decoding we used two test sets: to probe classifiers on the trained timbre (test *within*) we used the set of trials derived from the splitting procedure just mentioned. Within the same cross-validation loop, we assessed their ability to generalize on new spectral patterns (test *across*) by using all trials available for the untrained condition (for example, when the training set contained trials belonging to the cello sub-conditions, the test set across corresponded to all trials in which the viola was playing).
Chroma classification required two separate cross-validation rounds, differing in the composition of the sets employed. In the first round and within each loop, the training set included 90% of trials when either "C3" or "G4" were played; in this case, the test set corresponded to all trials belonging to the "C4" and "G3" conditions. Vice versa for the second round of cross-validation, 90% of the trials available for "C4" and "G3" were assigned to training and all the trials belonging to the alternative pair of notes were assigned to the test set. Again, cross-instrument decoding involved two test sets. For instance, when the training set was formed only by the sub-conditions "C3-cello" and "G4-cello", the test set *within* included "C4-cello" and "G3-cello" whereas the test set *across* was formed by "C4-viola" and "G3-viola".

Once the precise composition of training and test was established, we applied a "micro-averaging" procedure aimed at improving the signal-to-noise ratio in the data. At its core, this step consisted in forming pseudo-trials by averaging together groups of 8 epochs. Within this operation, to ensure an appropriate number of samples for learning, single epochs were used more than once. To maintain ad adequate and truthful level of variability in the data, essential to guarantee meaningful learning and fair tests, we made sure to minimize (a) the total number of times a given single epoch was used; and (b) the number of single epochs shared between any pair of the final pseudo-trials. Concerning training sets, groups to-be-averaged were defined with the aid of (constrained) permutations such that at least 18 pseudo-trials/note condition or 25 pseudo-trials/chroma condition were obtainable. In most cases, a single epoch was used 2 to 3 times and two pseudo-trials shared 2 to 3 single epochs at the maximum. Concerning the micro-averaging of the test sets, single epochs were used more than once only for height classification, with the exclusion of the test sets across instruments (in which, thanks to the independence between training and test data, many single trials were available). For height test sets, we formed 4 pseudo-trials per class where each pair shared 50% of single epochs at the maximum. Finally, when all sub-conditions were included in training (Figure 4.2), the amount of data available enabled us to use two micro-averaging alternatives: pseudo-trials composed of either 8 or 16 epochs. Whereas the improved signal-to-noise ratio of the latter version led to higher scores in absolute terms, these two alternatives yielded overlapping outcomes at the qualitative level.

Next, following the z-scoring each feature (i.e. channel and time point across trials), a L2-norm regularized Logistic Regression was fitted to the training set (Fan et al., 2008) in order to find the

hyperplane that could maximally predict *y* from *X* while minimizing a loss function. Since sometimes classes were slightly imbalanced in terms of number of pseudo-trials, a weighting procedure was always applied in order to equalize the contribution of each class to the definition of the hyperplane. The other model parameters were kept to their default values as provided by the Scikit-learn package.

After training, the models were used to predict *y* from the test set and the resulting (probabilistic) estimates were evaluated through comparison with the ground truth. More precisely, the probabilities estimated by both height and chroma classifiers were scored by computing the area under the Receiver Operating Characteristic curve (AUC), which summarizes the ratio between true and false positives. The value of AUC ranges between 0 and 1, with 0.5 corresponding to chance level. The scoring of height probabilities entailed a "one-vs-rest" scheme: AUC scores were computed for each class against the other three and then averaged. Additionally, striving for a more exhaustive evaluation, note estimates were further inspected by computing the corresponding error patterns. The latter were stored in confusion matrices where each entry *(i,j)* indicates the percentage of (pseudo-)trials belonging to class *i* and predicted as being part of *j*.

Lastly, the scores and confusion matrices obtained from all cross-validation runs were averaged within subject before any further step (e.g. group-level statistics).

For the main decoding analyses (i.e. when training incorporated all sub-conditions indistinctively, Figure 4.2 and 4.6) and within the same cross-validation procedure just described, classifiers trained on a given time lag *t* were tested not only at *t* but also at every other time lag *t'* composing the trial (King & Dehaene, 2014). The outcome of this procedure is a temporal generalization matrix where training times are ordered along the horizontal axis and rows display the performances obtained all along the trial. The inspection of this matrix enables to assess the similarity of the coding patterns as a function of time. The ability of a classifier to perform above-chance at multiple time points is indicative of the maintenance or re-occurrence of an informative neural activity pattern.

4.2.4. Multiple regression analysis on neural confusability

The first step for this complementary analysis consisted in training series of estimators on an 8-class problem where each spectral profile composing the auditory space was kept separate from the others (i.e. "C3-cello" vs "C3-viola" vs "G3-cello" vs "G3-viola" and so forth). We adopted a "one-vs-rest" approach and the same techniques described above. Within each cross-validation loop, we stored the error matrix displayed by the estimators at test, reporting the percentage of times each sub-condition was either correctly classified or mistaken for any of the other classes (Figure 4.4A).

Meanwhile, we built four theoretical matrices depicting stimuli's pairwise difference along four dimensions: fundamental frequency, center of gravity, spectral standard deviation and chroma. The first three parameters, assessed with the aid of Praat software (www.praat.org), are reported in Table 4.1. The center of gravity (also referred to as "centroid frequency" throughout the text) consists in the mean spectral frequency weighted by the distribution of signal amplitudes across the entire

spectrum and corresponds to an approximation of the overall frequency content of the stimulus. The standard deviation indicates how far the frequencies in the spectrum deviate from the center of gravity; we used this parameter as a measure of overall spectral shape. The distance, on a logarithmic scale, between each pair of stimuli along these three dimensions was then used to create the corresponding theoretical models. Logarithmic distances were preferred to plain differences to take into account that the discriminability of continuous quantitative parameters is known to follow Weber's law. The fourth theoretical matrix consisted of a categorical model in which two stimuli were entered as identical (distance=0) when they shared the same chroma and completely different (distance=1) in case they did not.

The core of the analysis consisted in fitting a multiple linear regression to ask whether the four theoretical matrices could explain the error patterns shown by the classifiers at each time point. All matrices were z-transformed before estimating the coefficients. Importantly, thanks to the inclusion of two musical instruments in the experimental stimulation, predictors were satisfactorily decorrelated one from the other (correlation between: F0 and spectral centroid = 0.28; F0 and spectral SD= 0.2; F0 and chroma=0.29; spectral centroid and SD = 0.35; spectral centroid and chroma = 0.29; spectral SD and chroma= 0.07) and the variance inflation factor for each of them was adequately low (1.17, 1.28, 1.16 and 1.15 from F0, spectral centroid, SD and chroma respectively). Given such prerequisites, the beta-weights obtained with this multiple regression reflect the specific and unique influence exerted by each variable on the patterns of neural confusion over and beyond the contribution of the remaining three. Thanks to this feature, we had the opportunity to (a) query whether the fundamental frequency of the sound is encoded separately from the other spectral components (b) test the authenticity of chroma encoding i.e. its separability from mere frequency-related processing.

As a note, two stimuli that are represented similarly at the neural level will give rise to a *high* percentage of misclassifications whereas a great degree of similarity along any of the predictor variables will be reflected by a *low* distance value. Thus, if any of our theoretical matrices captured a dimension that is actually encoded at the neural level we expected to find an effect with a negative sign (i.e. beta-weights that are significantly lower than zero).

4.2.5. Statistical analysis

Statistical analyses consisted of second-level tests across subjects and were implemented with MNE dedicated functions. Specifically, we used one-sample cluster-based permutation t-tests (Maris & Oostenveld, 2007), which intrinsically account for multiple comparisons, to determine whether (a) time-resolved classification scores were higher than chance; (b) decoding performances within the trained instrument were superior to those achieved across new spectra; and (c) multiple regression beta-weights were lower than zero. The analyses considered bidimensional clusters for the classification procedures producing temporal generalization matrices (shape: training times x testing times) and one-dimensional clusters otherwise. Univariate t-values were calculated for every

score/beta-weight with the exclusion of those corresponding to the baseline period. All samples that passed a threshold corresponding to a p-value of 0.05 were then grouped into clusters based on temporal adjacency. Cluster-level test statistics corresponded to the sum of t-values within each cluster. Their significance was computed by means of the Monte-Carlo method: they were compared to a null distribution of test statistics created by drawing 10000 random sign flips of the observed outcomes. A cluster was considered as significant when its p-value was below 0.05.

4.3. RESULTS

All the analyses presented below involved the employment of series of linear classifiers trained on brief (10ms), consecutive time-windows all along the event-related potentials (ERPs).

4.3.1. Decoding notes at sequence onset

The first procedure in our investigation consisted in asking whether (and when) infant neural responses could be classified according to the notes subjects were exposed to, each corresponding to a specific pitch height condition (C3=130.8Hz vs G3=196Hz vs C4=261.6Hz vs G4=392Hz). We observed that the identity of the note could be reliably discerned from the neural signal in between 130 and 680ms (Figure 4.2A-left), with the highest score obtained at 210ms relatively to the onset of the first tone (N=26; AUC=0.554±0.048, chance=0.5). A visual examination of the confusion matrices derived from these same classifiers (Figure 4.2A-center) suggests that, around the time of peak performance, each of the four notes could be satisfactorily distinguished from the other three and no systematic error occurred ("diagonal" shape). Still, we noticed a moderate increase in accuracy as a function of height with the best score observed for G4, the highest pitch condition. This phenomenon could be read as consistent with a more systematic assessment performed on adults with MEG (Krumbholz et al., 2003), showing how height modulates the amplitude of pitch-onset responses (i.e. higher frequencies elicit responses with greater amplitude).

To deepen our characterization of the temporal dynamics underlying decodability, classifiers were tested not only at the trained time window (Figure 4.2A-left) but also at every other time point along the trial. Since each classifier is specific to a given pattern of neural activity, assessing its performance across time enables to unravel the latency, duration and potential re-occurrence of the neural codes underlying successful classification (King & Dehaene, 2014). The matrix in Figure 4.2A (right) shows that successful height decoders could achieve above-chance performance at multiple time lags, giving rise to a diffuse generalization pattern with a mixed profile. First, a diagonal shape indicates the unfolding of an encoding *process* rather than a sustained time-invariant code. Further, we observed that early training led to above-chance performance up to ~680ms after sequence onset while classifiers trained later in the trial were able to generalize backward in time (i.e. from 120ms onwards). The resulting semi-squared shape indicates the presence of a metastable or repeating neural code, which supported note classification over multiple windows. The temporal dynamic just

described might derive from either the maintenance of note-related information, the re-activation of the same information by consecutive tones or, most likely, by a mixture of the latter two phenomena. Interestingly, when measured with MEG/EEG, the response elicited by pitch-producing sounds in adults is formed by both a transient spiky complex and a sustained component that remains steady for a prolonged period (Gutschalk et al., 2004). Similar neural dynamics in infants would account well for the decodability patterns observed: the peak in note classification scores reached at the very beginning of the trial (Figure 4.2A-left) as well as the diagonal pattern that arises for the matrix (Figure 4.2A-right) are consistent with the presence of a transient and composite phase. The sustained portion of the response (observed in adults), together with the re-activation of some of the initial components on the behalf of successive sounds is likely to give rise to the protracted periods of temporal generalization.

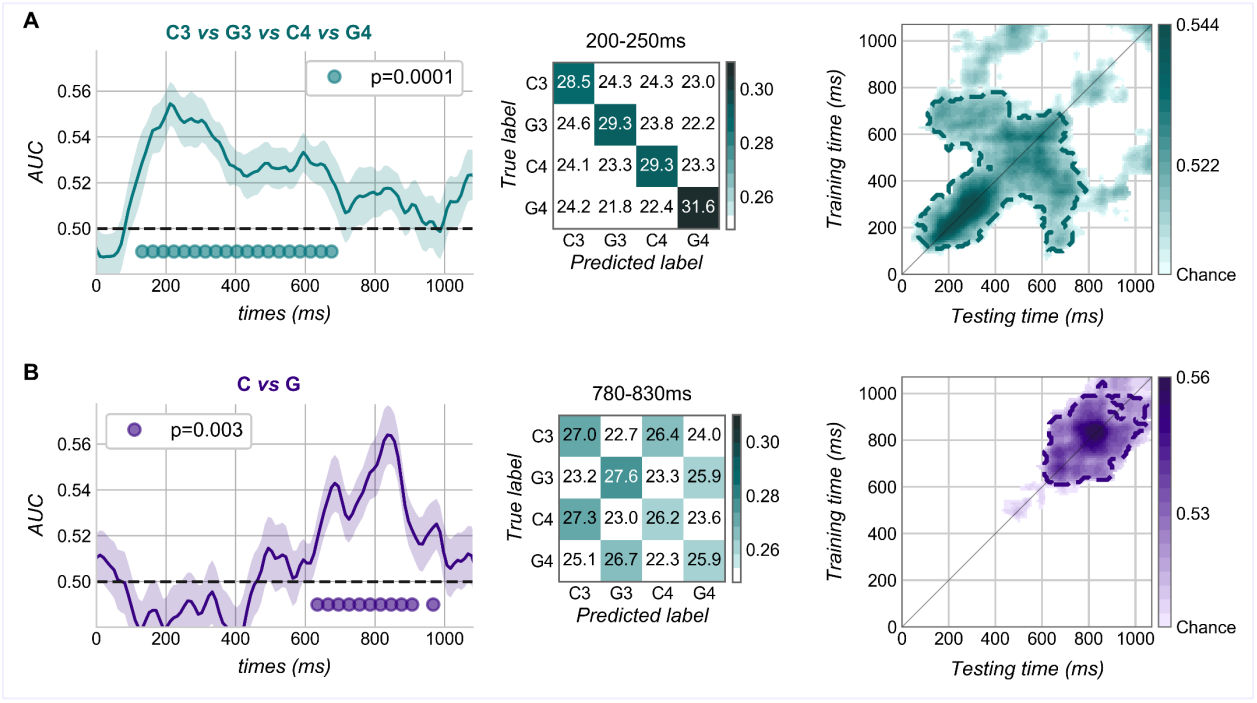

Figure 4.2 Average classification performances of estimators trained on single 10ms-windows from sequence onset. *(left)* Decoders are tested at the trained time sample. The shaded areas correspond to the SE (SEM) across subjects (N=26), the dotted black lines mark theoretical chance level. *(center)* Confusion matrices. The numbers within each cell indicate the percentage of times a given note from the y-axis was classified with the label reported on the x-axis. Off-diagonal values diverging from 0 signal misidentification (chance = 25%). Proportions below and at chance level are in white, every color step corresponds to .05%. Five time points, centered around peak AUC scores are averaged. *(right)* Temporal generalization matrices: above-chance AUC scores observed when performance is assessed at each consecutive time-point along the ERP. The diagonal thin lines demark classifiers trained and tested on the same time sample. The dashed contours indicate significant clusters (cluster-based permutation t-tests, the p-value is reported on the right).

Next, we asked whether trials could be reliably classified according to chroma distinctions (C vs G), irrespective of the particular frequency of the tones (i.e. independently from their height). This goal required a strategic training-test schema: separate sets of decoders were trained on trials when either C3 vs G4 or C4 vs G3 where presented and tested on the left-out pair of conditions. Such an expedient ensured that a height-based rule could not lead to above chance classification, since class "C" relatively to class "G" was characterized by a higher frequency during training and a lower frequency at test (or vice-versa). This decoding strategy yielded above-chance scores in between 630 and 990ms after sequence onset (Figure 4.2B-left), with a peak in performance observed at 830ms (AUC=0.564±0.092). The errors produced by note classifiers around this time were fully consistent with such an outcome: notes sharing the same chroma were systematically mistaken one for the other, giving rise to a "checkerboard" pattern of confusion (Figure 4.2B-center). When tested across time, classifiers successful in distinguishing C from G trials revealed a neatly distinct dynamic relatively to that observed for note identity: they generalized over a compact and quite restricted window (250ms-long on average), indicating the underlying presence of a unitary and transient neural code (Figure 4.2B-right).

4.3.2. Delineating the nature of the neural codes

The decoding performances collected so far demonstrate that infant brain responses contain explicit information related to both note identity and chroma. Yet, the nature of such information remains undetermined. In note classification, decodability might reflect the encoding of the fundamental frequencies as well as that of other spectral elements characterizing each note. For instance, independent codes for each of the eight spectral centroids included in the auditory space (Table 4.1) might suffice for classifiers to sort trials in four arbitrary groups. Crucially, as explained in the introduction, successful chroma decoding might simply derive from the physical similarity between pairs of tones along the frequency dimension.

To solve these ambiguities, we relied on cross-instrument classification. Namely, we tested the capability of estimators trained on a single timbre, i.e. a specific set of spectra, to transfer their learning at the alternative instrument, i.e. on a new set of spectra. Since according to a previous EEG investigation neonates can recognize musical notes across variations in resonator size (Háden et al., 2009), we expected to obtain successful cross-condition decoding for at least one set of estimators. In this analysis the number of trials available for training was markedly lower (half) relatively to before, such that a direct comparison with the previous outcomes might have been misleading. Thus, we used the scores achieved *within* context (i.e. on left-out trials belonging to the trained condition) as a benchmark to evaluate the performance observed *across* the new spectra.

When trained exclusively on either the cello or the viola condition, classifiers could successfully estimate the notes played by the alternative instrument in between 130 and 500ms. Figure 4.3A (left) shows how the generalization performance observed during this time window overlapped almost perfectly with the scores obtainable within the trained context. Also the confusion matrices, yielded

by these same classifiers around the peaks, were similar within and across context (Figure 4.3A-right): in both cases they revealed a diagonal error pattern, indicating that generalization occurred for all the notes, and a slightly more accurate performance within the fourth octave. Since the only feature shared by the two alternative training sets was the fundamental frequency characterizing each class, these results indicate that, during the first portion of the trial, the infant brain encoded the fundamental frequency of the sounds irrespective of their harmonics. Although our analysis did not highlight significant differences at any point along the trial, tests within the trained context relatively to those across context (and consistently with the main performance displayed in Figure 4.2A) produced significant scores for a protected period ranging from 500 to 680ms after sequence onset. Thus, it possible that after estimating the fundamental frequency the infant brain proceeded with the evaluation of other spectral features.

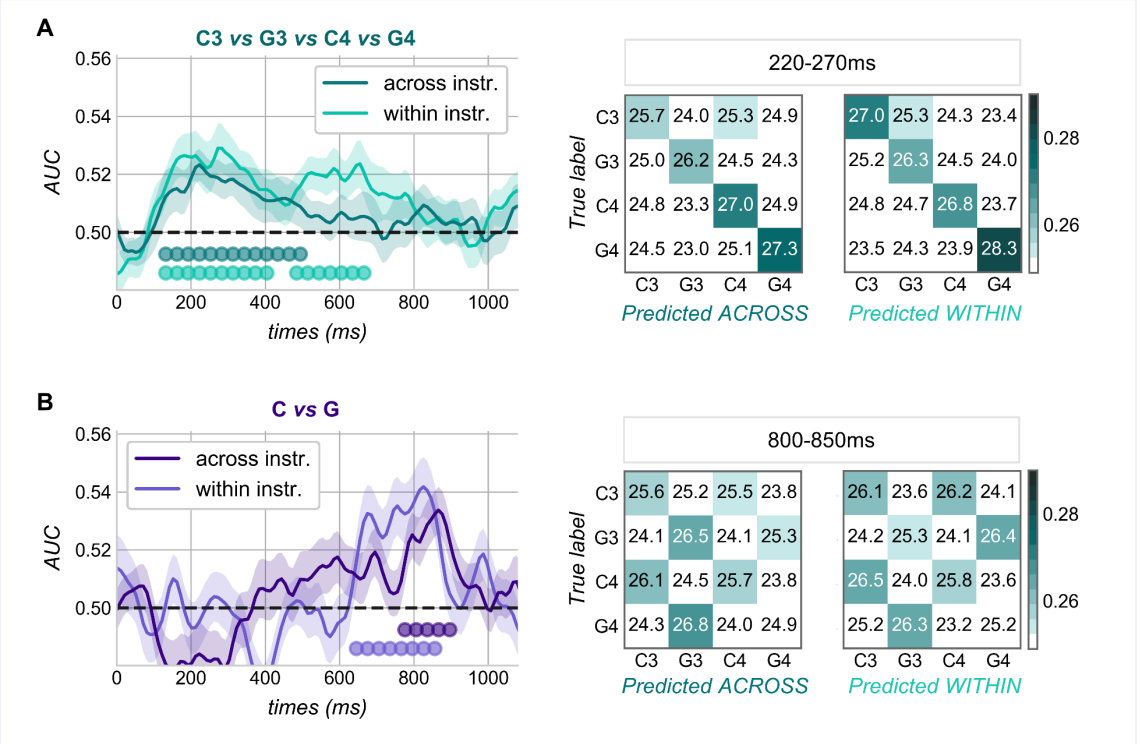

Figure 4.3 Cross-instrument decoding. *(left)* Time-resolved performance of estimators trained on a single musical instrument (e.g. note classifiers trained on cello). In light colors: classification within the trained condition (e.g. test on cello); in darker colors: performance at novel spectral configurations (e.g. test on viola). The scores from all possible training conditions or train/test directions are averaged. The shaded areas correspond to the SEM across subjects. Filled circles indicate above-chance scores (note classification within instrument: $p_{clust}<0.004$ and across: $p_{clust}=0.0001$; chroma classification within instrument: $p_{clust}=0.0168$ and across $p_{clust}=0.0006$) *(right)* Confusion matrices yielded by note classifiers at familiar and novel spectral contexts. As before, the numbers within each cell indicate the percentage of times a given note from the y-axis was (mis)classified with the label on the x-axis. Proportions below and at chance level are in white, each color step corresponds to .033%. Every matrix illustrates an average over 50ms (five time-points, i.e. five estimators) selected in between peak AUC scores.

Remarkably, classifiers trained on chroma distinctions at a single spectral context could generalize to the new timbre from 770 to 920ms after sequence onset (Figure 4.3B-left). The scores obtained across instruments were comparable to those achieved within the trained condition. Resembling the checkerboard pattern retrieved before and in full congruency with such a decoding outcome, note classifiers tested both within and across instruments during this late period showed a systematic mislabeling of C3/C4 and G3/G4 (Figure 4.3B-right). Overall, these results provide strong evidence for a genuine encoding of chroma on the behalf of the infant brain.

The examination of pairwise neural confusability validates cross-decoding outcomes

To provide additional evidence in support of our conclusions, we trained classifiers on eight separate classes, each corresponding to a spectral sub-condition (e.g. "C3-cello"), and used their errors as a proxy of pairwise similarity between neural responses. The misclassification patterns arising from this relatively un-supervised decoding scheme (Figure 4.4A-left) reveal which type of information is available for estimators to separate sub-conditions, thereby providing the possibility to discern which characteristics of the stimuli are encoded by the infant brain at a given moment. With a multiple linear regression, we tested whether (and when) neural confusion patterns could be explained by stimuli's differences along four dimensions: fundamental frequency, spectral centroid, spectral shape and chroma (Figure 4.4A-right). Given that all predictors were entered together in the regression, significant beta-weights assigned to a particular descriptor indicate that, at a given time point, the classifier relied on that particular dimension over and beyond the remaining three.

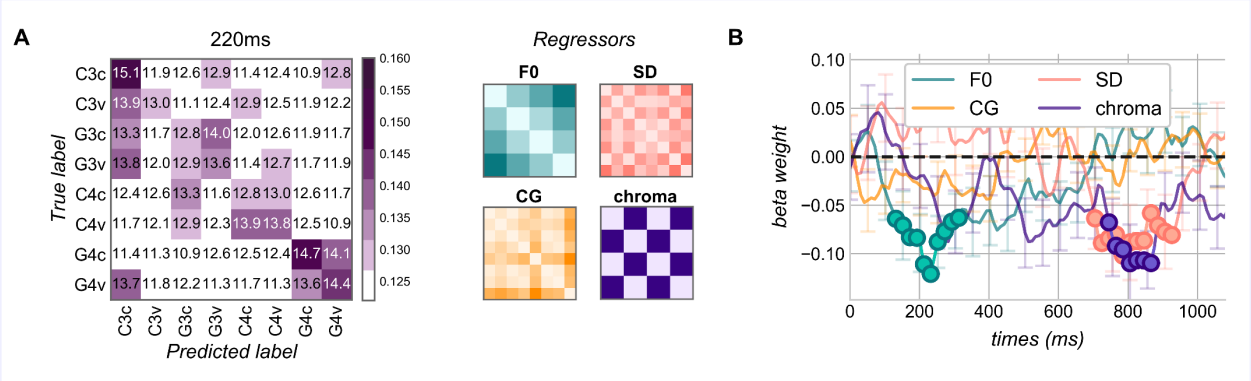

Figure 4.4 Multiple linear regression on neural confusability. (A) *Right* - example of confusion matrix produced by classifiers trained (at 220ms) on the eight classes specified by the axis labels ("c"=cello, "v"=viola). Numbers within each cell indicate the percentage of times a given spectral profile (i.e. sub-condition) was (mis)classified as indicated by x-axis (chance = 12.5%). Higher percentages suggest greater neural similarity. *Left* - predictor matrices: the first three (F0=fundamental frequency; CG=center of gravity aka spectral centroid; SD= standard deviation aka spectral shape) depict the distance on a logarithmic scale for the corresponding spectral dimension (Table 1); darker shades indicate greater dissimilarity. The fourth (chroma) is a categorical matrix where 1 stands for "different" (i.e. maximal dissimilarity) and 0 indicates correspondence. The ordering of the labels is equal to that on the left. The four matrices were entered together in the regression to explain neural note confusion at each time point. (B) The obtained

beta weights averaged across subjects and marked by filled circles when significantly below zero (cluster-based permutation t-tests; F0: p_{clust}=0.0058, SD: p_{clust}=0.004, chroma: p_{clust}=0.0115). The vertical lines correspond to the SEM. A moving average with a window of two points was applied to enhance the clarity of the illustration.

Complementary to the decoding outcomes reported above, Figure 4.4B shows that differences in fundamental frequency predicted the pattern of neural separability during the first portion of the trial (130-320ms). Whereas the weights attributed to the centroid of the spectrum never reached significance, its shape drove neural confusability in between 700 and 900ms, indicating that, at least within this window, some spectral characteristic other than the F0 was processed. Crucially, at about the same time (740-880ms) chroma distinctions exerted a significant effect beyond that of merely spectral parameters. This latter outcome is noteworthy since it shows how a methodology that is parallel to cross-condition decoding provides the same evidence: a genuine encoding of chroma, not explicable in terms of physical (harmonic) overlap.

4.3.3. The time course of height and chroma processing

So far, we have demonstrated that, when exposed to auditory sequences, the infant brain processes the two psychophysical dimensions characterizing their pitch in a separate fashion: height decodability, reflecting the encoding of tone fundamental frequency, starts promptly right after the beginning of the stimuli whereas chroma is computed markedly later within the trial.

At this point, we queried how height and chroma processing unfolds when tones keep being repetitively played. As a matter of fact, the classification time courses observed within the previous analyses could have been molded by the presence of a conspicuous amount of sequences composed of four medium tones ("4M" condition): within these trials sound offset occurred at 660ms, a time-stamp roughly corresponding to the decay of height decodability and to the onset of successful chroma classification (Figure 4.2-left). To probe such an eventuality, we trained estimators exclusively on portions of the neural signal recorded while sounds kept being played. Importantly, to guarantee interpretability, this training set was balanced in terms of tone-durations/rate of presentation (i.e. training was conducted on an even number of "M" and "L" trials).

Figure 4.5A shows that height classifiers obtained significant scores over a single time-window (180-500ms): their performance declined at 600ms and remained at chance thereafter, despite the sounds were still ongoing. This result is compatible with a recent adult study reporting that when stimuli (iterated rippled noise segments, IRNs) with the same fundamental frequency are repeated, the auditory fields evoked by the sounds following the first are markedly reduced in amplitude (Andermann et al., 2021). In general, the adaptation effect observed here might have been exacerbated by the fact that sequence structure (e.g. the identity of the tones after the first) was fully predictable (Todorovic et al., 2011).

Conversely, chroma classifiers obtained above-chance scores not only during a first time-window (660-870ms), which corresponded to that observed in the initial analysis, but also within a second time window spanning from 1500 to 1690ms relatively to the onset of the first tone (Figure 4.5C). Thus, the infant brain kept tracking the quality of the sounds despite its predictability.

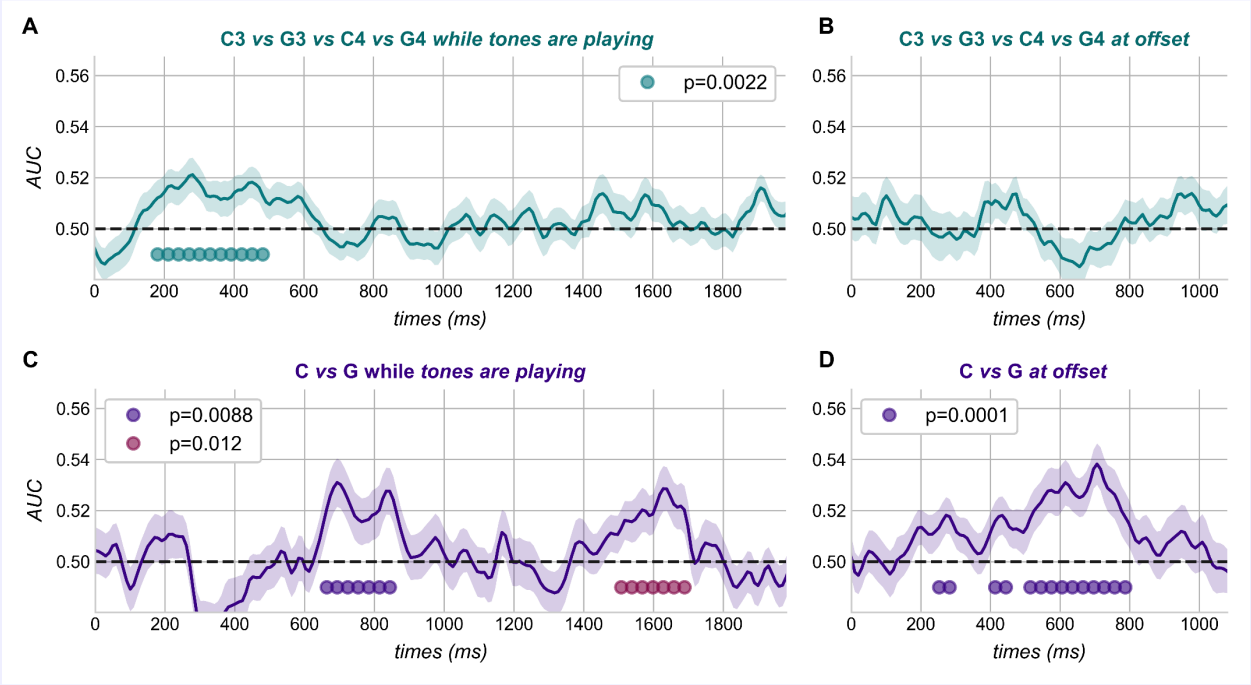

Figure 4.5 Time-course of pitch height and pitch chroma processing. *(left - A and C)* Classification performances observed throughout 2 seconds of continuous stimulation: the right extremity of the time axis coincides with the offset of either the 11th medium tone or the 4th long tone. Note that absolute scores are neatly reduced relatively to Figure 4.2 due to a markedly lower number of trials available for the analysis. *(right - B and D)* Classification performance during the silent period following the auditory sequence; zero corresponds to the onset of the last tone. In all graphs, shaded areas correspond to the SEM and dotted black lines mark theoretical chance level.

In the final portion of our investigation, we asked whether any pitch-specific process occurred when sounds vanished. With this aim, we applied height and chroma estimators to the neural signal recorded (during a silent period) at sequence offset. Figures 5B and 5D show the classification scores obtained when time 0 was set at the onset of the last note. As observed within sequence, the retrievability of height and chroma information diverged neatly. Concerning the former, the performance of note classifiers remained at chance throughout the entire (silent) period (Figure 4.5B). Such a null result deviates from the recent observations by Andermann and colleagues (2021), who report the presence of F0-related information in the MEG signal recorded at the offset of IRN segments. On the other end, the absence of height decodability at the end of the auditory sequences conforms the idea that, at least in adults, offset responses to pitch-evoking stimuli might be unrelated

to height itself but rather concern the cessation of the sound, the interruption of regularity or some other characteristic (Krishnan et al., 2014; Gutschalk et al., 2004).

Despite the impossibility to distinguish the precise identity of the notes, "C" and "G" trials could be reliably discerned at various time points during the silent window following the sequence and especially in between 510 and 790ms (Figure 4.5D). Resembling our previous finding (Figure 4.3B), decoders trained on single spectral conditions performed comparably within the same context and across the alternative instrument (Figure 4.6A), indicating that successful classification did not derive from shared harmonic components (i.e. genuine chroma decoding). At offset, the inspection of the performance across time revealed a much broader generalizability relative to what observed earlier in the trial (Figure 4.6B): for instance, classifiers trained in between 300 and 800ms could successfully transfer their learning across 45 time windows (corresponding to 450ms) on average. The diffuse and semi-squared pattern arising from the generalization matrix might suggest that the neural code for chroma was still unitary, as in the previous case, but maintained for longer in the absence of ongoing notes.

Given these results, we questioned whether either chroma decodability itself or its particular shape depended upon the number of tones contained in the auditory sequence. Since "4" and "12" trials were equally frequent within the experimental session, every 4^{th} tone was followed by silence only in 50% of cases while every 12^{th} tone corresponded to sequence offset in 100% of cases. Thus, it might be possible that chroma decodability was linked to some sort of expectation of a 5^{th} tone which never occurred. We reasoned that if that was the case, chroma-related information should be discernible only at the end of sequences composed of 4 tones, but not following 12 tones. Figure 4.6C shows the decoding performances observed when the analysis concerned only one numerosity condition at a time. Chroma-specific information was retrievable from the neural signal recorded at the offset of both types of sequences and in both cases successful classifiers could generalize to a broad range of time points. Yet, the latency of chroma decodability was different between the two conditions, with a shorter latency observed for trials in which 12 sounds were presented. The full interpretability of the latter observation is compromised by the fact that the two sets of trials did not differ only for numerosity but also for average sequence duration. Nevertheless, this assessment enabled us to ascertain that chroma-specific information was retrievable during silent periods irrespective of whether additional tones could be expected or not.[16]

Lastly, since the approaches found in the literature are heterogeneous in this regard, we performed the same decoding analysis in two alternative versions: trials were aligned upon either the actual offset of the last tone or the offset of the sequence (i.e. tone offset time + a silent gap corresponding to inter-tone-interval characterizing the trial). The outcomes obtained with these two alternative

[16] The performance of height classifiers remained at chance for both numerosity conditions.

trial alignments were equivalent to what described so far: classifiers succeeded at discerning chroma but not height.

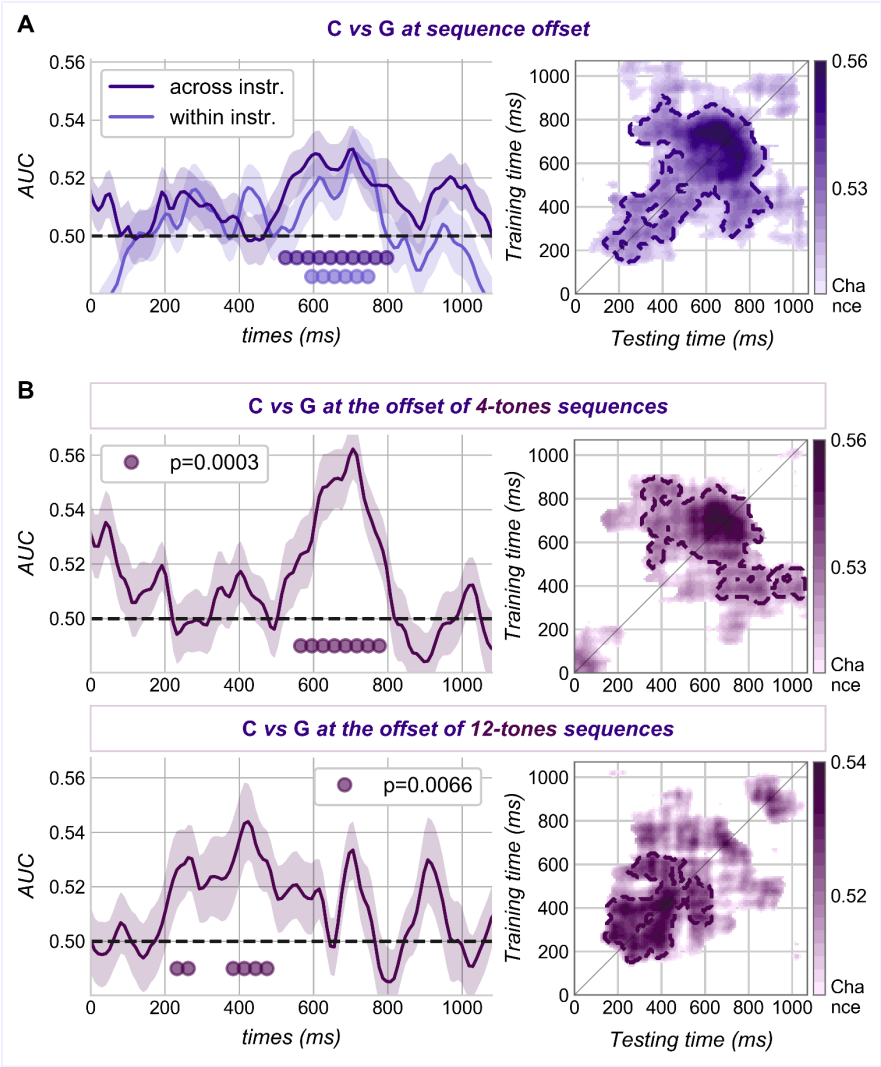

Figure 4.6 In-depth characterization of chroma decodability at sound offset

4.4. DISCUSSION

In the present study, we applied multivariate time-resolved decoding to high-resolution EEG recordings in order to characterize how the human brain encodes pitch early in life. Although related to a single physical factor (i.e. the repetition rate of the waveform), pitch has been described as composed of two perceptual dimensions, height and chroma, presupposing a counterpart for both components at the neural level. Meanwhile, an ongoing debate pertains to the idea that chroma, unlike height, might correspond to a higher-order and culturally determined construct which requires learning, attention and working memory in order to be represented. Our results

demonstrate that, when exposed to musical tones organized in tight and repetitive auditory streams, 3-mo-old infants encode both pitch height and pitch chroma in a completely automatic fashion. Strikingly, the classification performances we observed indicate that the two dimensions are processed with neatly divergent dynamics. Pitch height, as isolated by decoders trained on note identity, was computed soon after the onset of the stimuli, whereas the encoding of chroma began much later in the trial, at times when the decodability of the former declined. Further, only the neural processing of chroma persisted throughout the auditory stimulation (while tones were still being played) and after sound offset.

Drawing a parallelism with adult electrophysiology, our results relate to the findings reported by Briley and colleagues (2013) in various aspects. These authors were able to isolate the existence of a cortical mechanism specialized for chroma by showing that pitch-evoked activity adapts more to sounds at octave distance than to sounds at narrower musical intervals (Briley et al., 2013). In a way, such a demonstration mirrors the ability of our decoders to reliably classify "C" vs "G" despite the inconsistency of between-class height differences from training to test (i.e. the rule "high vs low frequency" would have led to scores below chance). In adults, the chroma effect was observed with unresolved IRN stimuli, which are characterized by a uniform distribution of frequency components, excluding the possibility for mere harmonic overlap to be at its origin. Equivalently, the ability of our decoders to generalize across new spectral profiles ensures a genuine code for chroma rather than physical similarity to be the basis of the decodability observed.
On the other hand, the retrieval of spontaneous and pre-attentive chroma processing[17], stands at odds with the results of Regev and coworkers (2019), who report the absence of mismatch cortical responses to chroma deviants in adults. We argue that such null finding might be related to the precise paradigm employed by these authors. Namely, whereas complex natural tones engage the cortex in meaningful and behaviorally relevant computations (Hoeschele et al., 2015), Regev and colleagues based their conclusions on pure tones, which have been repetitively pointed out as a class of stimuli lacking ecological relevance (Oxenham, 2018). For example, it is worth considering that pure tones have been proven inefficient in driving non-primary auditory neurons (Rauschecker et al., 1995; Wessinger et al., 2001). Alternatively, the paradigm in question may have actually triggered a chroma effect and simply lacked enough power: in the two experiments included in the study (Regev et al., 2019) changes in chroma did elicit a slow trend of differential activity that, however, did not reach statistical significance. Being discernible relatively late in the trial, the latency of such a trend in adults fits well with the temporal dynamics found here for infant chroma processing.

The divergence between the classification patterns observed for height and chroma is likely to originate from distinct substrate mechanisms serving different functional roles.

[17] the reader should keep in mind that infants slept for most of the experimental session

Auditory neurons are characterized by a preferred frequency to which they respond most strongly. Neurons with similar frequency preference cluster together, forming tonotopic maps that are found at various relays along the auditory hierarchy (Saenz & Langers, 2014), including primary and non-primary cortical regions (Formisano et al., 2003; Langers & van Dijk, 2012; Moerel et al., 2012). Such topographic organization is likely to constitute the basis for accurate (absolute) frequency estimations. In adults, the experience of sound height is modulated by both its fundamental frequency, corresponding to the strongest acoustical correlate of pitch height, and the center of gravity of the spectrum (or centroid frequency), which regulates the perceived brightness (Oxenham, 2013). To disentangle these two factors, each F0 included in our paradigm was coupled with two alternative spectral centroids, resulting in brighter notes for the viola condition. While estimators could have been driven by both parameters, our analyses (Figure 4.3A and Figure 4.4) testify that successful note classification relied on a neural code for the fundamental frequency. The functional role of such a code is reflected by the importance of the fundamental frequency in representing source identity and in segregating distinct sources into separate perceptual streams. For instance, F0 has been proven an effective cue when it comes to identify speakers (Van Dommelen, 1990; Baumann & Belin, 2008) and discern the boundaries of contiguous auditory events (Bregman, 1994). Coherently with our results, these capabilities seem to be available (to a certain extent) very early in development: since their first days of life, infants recognize their mother's voice (Spence & Freeman, 1996), discriminate unfamiliar speakers (Floccia et al., 2000) and make use of pitch cues to segregate simultaneously active sound sources (Winkler et al., 2003).

Beyond their preferred frequency, ~60% of the auditory neural populations within the human cortex respond to multiple additional frequency bands, resulting in a complex multi-peaked spectral tuning (Moerel et al., 2013). Of particular interest for our study, fMRI assessments on adults have revealed the existence (among others) of spatially distributed neuronal clusters that are specifically tuned to frequencies situated exactly one octave apart (Moerel et al., 2013, 2015). We interpret the ability of our classifiers to discern C from G trials as indicating that such a refined tuning to multiple octaves is in place already by the age of ~12 weeks. At the behavioral level, neuronal clusters with this characteristic are very likely to contribute to the perception of octave equivalence found not only in adults (Hoeschele et al., 2012) but also in 3 month-olds (Demany & Armand, 1984).

It has been proposed that neurons with a complex spectral tuning might serve to signal the presence of combinations of frequencies that are ecologically important (Kadia & Wang, 2003; Sadagopan & Wang, 2009). In this regard, what could be the functional role of chroma? We envision two (non-mutually exclusive) possibilities that could justify the presence of a dedicated neural mechanism so early in life. First, octave equivalence might play an important role in learning how to produce speech. Articulatory skills are acquired through vocal plays aimed at imitating ambient language: through trials and errors, infants progressively adjust their utterances to match those heard from their surroundings (Kuhl et al., 2008). Problematically, the voices that needs to be mimicked in this process are those of older humans, whose frequency range is too low for the young learner to replicate. In

this context, chroma comes into help: when reproducing the fundamental frequency is not feasible, transposing the sounds by one or more octaves results in the closest possible approximation (Hoeschele, 2017). Consistently with this proposition, it has been observed that when young children are asked to imitate non-words presented below their vocal range, they spontaneously reproduce the stimuli one octave above the target (Peter et al., 2008, 2009).

More broadly, while height indicates the source of the sound, chroma can be conceived as the holder of its message (Warren et al., 2003). That is, other than important cues to identify and localize emitters, pitch carries several other types of crucial information (for a review: Braun & Johnson, 2011). To start with, pitch movements define intonation, enabling us to discern questions from statements or to focus on certain portions of the sentence. Moreover, they convey attitudes and emotions, holding an inestimable communicative value in both music and speech. When it comes to track these types of information absolute frequency values are often unreliable due to spurious and prominent contextual variability. Chromatic features, on the other hand, are far more robust. For instance, if the cello and the viola used in our paradigm were performing a melody together within a real concert, they would certainly not use the same notes; instead, they would play notes with the same chroma. Correspondingly, the warning «do not touch the stove! » does not have the same fundamental frequencies when yelled by either mom or dad since, on average, (human) male and female voices stand one octave apart (Titze, 2000). Yet, precisely due the latter fact, mom and dad will produce sounds that have the same quality, thereby carrying overlapping meaning and emotional content[18] These scenarios emphasize how tracking chroma patterns might be an adaptive strategy, as it enables to form coherent information streams that become analyzable independently from their specific source[19]. Computationally, one way of implementing this strategy might correspond to auditory filters: according to psychoacoustic evidence, cuing a given frequency results in enhanced sensitivity to both the primed frequency and those placed one and two octaves above or below the cue (Borra et al., 2013).

As a conclusive consideration, we point out that the distinct roles proposed for height and chroma are coherent with the temporal progression of their decodability, as observed here. Namely, given the properties of our experimental stimulation (i.e. repetitive sounds from single emitters), it seems reasonable and efficient for the brain to isolate the source at first and follow (only) the message thereafter and for a protracted period.

[18] The statistical analysis of human natural speech reveals that the frequency ratios defining the structure of the chromatic scale (which is universal in tonal music) correspond to the empirical concentrations of power in linguistic sounds (Schwartz et al., 2003). Thus, the two example just described appear intimately interconnected; an observation that highlights the relevance of our findings.

[19] According to a recent study, this may be precisely what adults do. Namely, when height and chroma cues are incongruent, subjects have the spontaneous tendency to rely systematically on chroma when asked to judge the direction of pitch shifts (Lin et al., 2018).

To summarize, previous electrophysiological investigations had revealed that "proto-pitch" features are represented since birth (Háden et al., 2009) and further suggested that a refined "adult-like" processing of pitch starts to appear only by the age of 4 months (He & Trainor, 2009). Regardless of such immaturity, our study shows how younger infants encode pitch along two separate dimensions: height and chroma. Since exposure to tonal stimuli (e.g. speech and music) begins in utero, our results do not allow strong claims in favor of biologically-determined mechanisms. Still, they demonstrate that pitch chroma is not a high-order construct but rather a basic organizing principle of neural responsivity observable very early in development.

4.5. References

Andermann, M., Günther, M., Patterson, R. D., & Rupp, A. (2021). Early cortical processing of pitch height and the role of adaptation and musicality. *NeuroImage*, *225*, 117501. https://doi.org/10.1016/j.neuroimage.2020.117501

Apel, W. (2003). *The Harvard Dictionary of Music: Fourth Edition*. Harvard University Press.

Baumann, O., & Belin, P. (2008). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research PRPF*, *74*(1), 110. https://doi.org/10.1007/s00426-008-0185-z

Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature*, *436*(7054), 1161–1165. https://doi.org/10.1038/nature03867

Bitterman, Y., Mukamel, R., Malach, R., Fried, I., & Nelken, I. (2008). Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature*, *451*(7175), 197–201. https://doi.org/10.1038/nature06476

Borra, T., Versnel, H., Kemner, C., van Opstal, A. J., & van Ee, R. (2013). Octave effect in auditory attention. *Proceedings of the National Academy of Sciences*, *110*(38), 15225–15230. https://doi.org/10.1073/pnas.1213756110

Braun, B., & Johnson, E. K. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *Journal of Phonetics*, *39*(4), 585–594. https://doi.org/10.1016/j.wocn.2011.06.002

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

Briley, P. M., Breakey, C., & Krumbholz, K. (2013). Evidence for Pitch Chroma Mapping in Human Auditory Cortex. *Cerebral Cortex*, *23*(11), 2601–2610. https://doi.org/10.1093/cercor/bhs242

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings 17, 97–110.

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer* (Version 6.0.25) [Computer software]. http://www.praat.org/

Burns, E. M. (1999). 7—Intervals, Scales, and Tuning**This chapter is dedicated to the memory of W. Dixon Ward. In D. Deutsch (Ed.), *The Psychology of Music (Second Edition)* (pp. 215–264). Academic Press. https://doi.org/10.1016/B978-012213564-4/50008-1

Cansino, S., Ducorps, A., & Ragot, R. (2003). Tonotopic cortical representation of periodic complex sounds. *Human Brain Mapping*, *20*(2), 71–81. https://doi.org/10.1002/hbm.10132

Crottaz-Herbette, S., & Ragot, R. (2000). Perception of complex sounds: N1 latency codes pitch and topography codes spectra. *Clinical Neurophysiology*, *111*(10), 1759–1766. https://doi.org/10.1016/S1388-2457(00)00422-3

Curtis, M. E., & Bharucha, J. J. (2010). The minor third communicates sadness in speech, mirroring its use in music. *Emotion*, *10*(3), 335–348. https://doi.org/10.1037/a0017928

Demany, L., & Armand, F. (1984). The perceptual reality of tone chroma in early infancy. *The Journal of the Acoustical Society of America*, *76*(1), 57–66. https://doi.org/10.1121/1.391006

Dinther, R. van, & Patterson, R. D. (2006). Perception of acoustic scale and size in musical instrument sounds. The Journal of the Acoustical Society of America, 120(4), 2158–2176. https://doi.org/10.1121/1.2338295

Draganova, R., Eswaran, H., Murphy, P., Huotilainen, M., Lowery, C., & Preissl, H. (2005). Sound frequency change detection in fetuses and newborns, a magnetoencephalographic study. *NeuroImage*, *28*(2), 354–361. https://doi.org/10.1016/j.neuroimage.2005.06.011

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*(Aug), 1871–1874.

Fernald, A. (1989). Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message? *Child Development*, *60*(6), 1497–1510. https://doi.org/10.2307/1130938

Floccia, C., Nazzi, T., & Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns. *Developmental Science*, *3*(3), 333–343. https://doi.org/10.1111/1467-7687.00128

Formisano, E., Kim, D.-S., Di Salle, F., van de Moortele, P.-F., Ugurbil, K., & Goebel, R. (2003). Mirror-Symmetric Tonotopic Maps in Human Primary Auditory Cortex. *Neuron*, *40*(4), 859–869. https://doi.org/10.1016/S0896-6273(03)00669-X

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

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

Gutschalk, A., Patterson, R. D., Scherg, M., Uppenkamp, S., & Rupp, A. (2004). Temporal dynamics of pitch in human auditory cortex. *NeuroImage*, *22*(2), 755–766. https://doi.org/10.1016/j.neuroimage.2004.01.025

Háden, G. P., Stefanics, G., Vestergaard, M. D., Denham, S. L., Sziller, I., & Winkler, I. (2009). Timbre-independent extraction of pitch in newborn infants. *Psychophysiology*, *46*(1), 69–74. https://doi.org/10.1111/j.1469-8986.2008.00749.x

He, C., Hotson, L., & Trainor, L. J. (2007). Mismatch Responses to Pitch Changes in Early Infancy. *Journal of Cognitive Neuroscience*, *19*(5), 878–892. https://doi.org/10.1162/jocn.2007.19.5.878

He, C., & Trainor, L. J. (2009). Finding the Pitch of the Missing Fundamental in Infants. *Journal of Neuroscience*, *29*(24), 7718–8822. https://doi.org/10.1523/JNEUROSCI.0157-09.2009

Hoeschele, M. (2017). Animal Pitch Perception: Melodies and Harmonies. *Comparative Cognition & Behavior Reviews*, *12*. http://comparative-cognition-and-behavior-reviews.org/2017/vol12_hoeschele/

Hoeschele, M., Merchant, H., Kikuchi, Y., Hattori, Y., & ten Cate, C. (2015). Searching for the origins of musicality across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140094. https://doi.org/10.1098/rstb.2014.0094

Hoeschele, M., Weisman, R. G., & Sturdy, C. B. (2012). Pitch chroma discrimination, generalization, and transfer tests of octave equivalence in humans. *Attention, Perception, & Psychophysics*, *74*(8), 1742–1760. https://doi.org/10.3758/s13414-012-0364-2

Jacoby, N., Undurraga, E. A., McPherson, M. J., Valdés, J., Ossandón, T., & McDermott, J. H. (2019). Universal and Non-universal Features of Musical Pitch Perception Revealed by Singing. *Current Biology*, *29*(19), 3229-3243.e12. https://doi.org/10.1016/j.cub.2019.08.020

Kadia, S. C., & Wang, X. (2003). Spectral Integration in A1 of Awake Primates: Neurons With Single- and Multipeaked Tuning Characteristics. *Journal of Neurophysiology*, *89*(3), 1603–1622. https://doi.org/10.1152/jn.00271.2001

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Krishnan, A., Gandour, J. T., Ananthakrishnan, S., & Vijayaraghavan, V. (2014). Cortical pitch response components index stimulus onset/offset and dynamic features of pitch contours. *Neuropsychologia*, *59*, 1–12. https://doi.org/10.1016/j.neuropsychologia.2014.04.006

Krumbholz, K., Patterson, R. D., Seither-Preisler, A., Lammertmann, C., & Lütkenhöner, B. (2003). Neuromagnetic Evidence for a Pitch Processing Center in Heschl's Gyrus. *Cerebral Cortex*, *13*(7), 765–772. https://doi.org/10.1093/cercor/13.7.765

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Langers, D. R. M., & van Dijk, P. (2012). Mapping the Tonotopic Organization in Human Auditory Cortex with Minimally Salient Acoustic Stimulation. *Cerebral Cortex*, *22*(9), 2024–2038. https://doi.org/10.1093/cercor/bhr282

Lau, B. K., & Werner, L. A. (2012). Perception of missing fundamental pitch by 3- and 4-month-old human infants. *The Journal of the Acoustical Society of America*, *132*(6), 3874–3882. https://doi.org/10.1121/1.4763991

Lecanuet, J. P., Graniere-Deferre, C., Jacquet, A.-Y., & DeCasper, A. J. (2000). Fetal discrimination of low-pitched musical notes. *Developmental Psychobiology*, *36*(1), 29–39. https://doi.org/10.1002/(SICI)1098-2302(200001)36:1<29::AID-DEV4>3.0.CO;2-J

Lin, J.-F. L., Imada, T., Kuhl, P. K., & Lin, F.-H. (2018). Incongruent pitch cues are associated with increased activation and functional connectivity in the frontal areas. *Scientific Reports*, *8*(1), 5206. https://doi.org/10.1038/s41598-018-23287-5

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

Moerel, M., De Martino, F., Santoro, R., Yacoub, E., & Formisano, E. (2015). Representation of pitch chroma by multi-peak spectral tuning in human auditory cortex. *NeuroImage*, *106*, 161–169. https://doi.org/10.1016/j.neuroimage.2014.11.044

Moerel, M., Martino, F. D., & Formisano, E. (2012). Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *Journal of Neuroscience*, *32*(41), 14205–14216. https://doi.org/10.1523/JNEUROSCI.1388-12.2012

Moerel, M., Martino, F. D., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., & Formisano, E. (2013). Processing of Natural Sounds: Characterization of Multipeak Spectral Tuning in Human Auditory Cortex. *Journal of Neuroscience*, *33*(29), 11888–11898. https://doi.org/10.1523/JNEUROSCI.5306-12.2013

Oxenham, A. J. (2013). 1—The Perception of Musical Tones. In D. Deutsch (Ed.), *The Psychology of Music (Third Edition)* (pp. 1–33). Academic Press. https://doi.org/10.1016/B978-0-12-381460-9.00001-8

Oxenham, A. J. (2018). How We Hear: The Perception and Neural Coding of Sound. *Annual Review of Psychology*, *69*(1), 27–50. https://doi.org/10.1146/annurev-psych-122216-011635

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Peter, B., Larkin, T., & Stoel-Gammon, C. (2009). Octave-shifted pitch matching in nonword imitations: The effects of lexical stress and speech sound disorder. *The Journal of the Acoustical Society of America*, *126*(4), 1663–1666. https://doi.org/10.1121/1.3203993

Peter, B., Stoel-Gammon, C., & Kim, D. (2008). Octave equivalence as an aspect of stimulus-response similarity during nonword and sentence imitations in young children. *Proceedings of the 4th International Conference on Speech Prosody, SP 2008*, 731–734. https://asu.pure.elsevier.com/en/publications/octave-equivalence-as-an-aspect-of-stimulus-response-similarity-d

Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, *268*(5207), 111–114. https://doi.org/10.1126/science.7701330

Regev, T. I., Nelken, I., & Deouell, L. Y. (2019). Evidence for Linear but Not Helical Automatic Representation of Pitch in the Human Auditory System. *Journal of Cognitive Neuroscience*, *31*(5), 669–685. https://doi.org/10.1162/jocn_a_01374

Sadagopan, S., & Wang, X. (2009). Nonlinear Spectrotemporal Interactions Underlying Selectivity for Complex Sounds in Auditory Cortex. *Journal of Neuroscience*, *29*(36), 11192–11202. https://doi.org/10.1523/JNEUROSCI.1286-09.2009

Saenz, M., & Langers, D. R. M. (2014). Tonotopic mapping of human auditory cortex. *Hearing Research*, *307*, 42–52. https://doi.org/10.1016/j.heares.2013.07.016

Schellenberg, E. G., & Trehub, S. E. (1996). Natural Musical Intervals: Evidence From Infant Listeners. *Psychological Science*, *7*(5), 272–277. https://doi.org/10.1111/j.1467-9280.1996.tb00373.x

Schwartz, D. A., Howe, C. Q., & Purves, D. (2003). The Statistical Structure of Human Speech Sounds Predicts Musical Universals. *The Journal of Neuroscience*, *23*(18), 7160–7168. https://doi.org/10.1523/JNEUROSCI.23-18-07160.2003

Sergeant, D. (1983). The Octave-Percept or Concept. *Psychology of Music*, *11*(1), 3–18. https://doi.org/10.1177/0305735683111001

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, *89*(4), 305–333. https://doi.org/10.1037/0033-295X.89.4.305

Spence, M. J., & Freeman, M. S. (1996). Newborn infants prefer the maternal low-pass filtered voice, but not the maternal whispered voice. *Infant Behavior and Development*, *19*(2), 199–212. https://doi.org/10.1016/S0163-6383(96)90019-3

Titze, I. R. (2000). *Principles of voice production (second printing).* IA: National Center for Voice and Speech.

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. *Journal of Neuroscience*, *31*(25), 9118–9123. https://doi.org/10.1523/JNEUROSCI.1425-11.2011

Trainor, L. J., Tsang, C. D., & Cheung, V. H. W. (2002). Preference for Sensory Consonance in 2- and 4-Month-Old Infants. *Music Perception: An Interdisciplinary Journal*, *20*(2), 187–194. https://doi.org/10.1525/mp.2002.20.2.187

Van Dommelen, W. A. (1990). Acoustic Parameters in Human Speaker Recognition. *Language and Speech*, *33*(3), 259–272. https://doi.org/10.1177/002383099003300302

Warren, J. D., Jennings, A. R., & Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *NeuroImage*, *24*(4), 1052–1057. https://doi.org/10.1016/j.neuroimage.2004.10.031

Warren, J. D., Uppenkamp, S., Patterson, R. D., & Griffiths, T. D. (2003). Separating pitch chroma and pitch height in the human brain. *Proceedings of the National Academy of Sciences*, *100*(17), 10038–10042. https://doi.org/10.1073/pnas.1730682100

Werner, L., Fay, R. R., & Popper, A. N. (Eds.). (2012). *Human Auditory Development* (Vol. 42). Springer New York. https://doi.org/10.1007/978-1-4614-1421-6

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical Organization of the Human Auditory Cortex Revealed by Functional Magnetic Resonance Imaging. *Journal of Cognitive Neuroscience*, *13*(1), 1–7. https://doi.org/10.1162/089892901564108

Winkler, I., Kushnerenko, E., Horvath, J., Ceponiene, R., Fellman, V., Huotilainen, M., Naatanen, R., & Sussman, E. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences*, *100*(20), 11812–11815. https://doi.org/10.1073/pnas.2031891100

Chapter 5. GENERAL DISCUSSION & PERSPECTIVES

5.1. A common line between long-lasting debates

With the experimental work described throughout the present thesis, we sought to reveal the encoding processes engaged by the infant brain within three perceptual domains: speech, numerical quantity and pitch. Our choice was motivated by both the centrality of these domains in human every-day life, and the aspiration to solve long-lasting debates which exist in regard to each one of them.
Whereas some investigators point out that phonetic processing might be available since start (Dehaene-Lambertz, 2017), others dismiss such an eventuality favoring the early employment of domain-general acoustic principles for speech analysis (Kuhl et al., 2008). Even throughout the literature on adults, there is no agreement upon the nature of the basic units serving speech perception as we are used to experience it (Frauenfelder & Floccia, 1998; Kazanina et al., 2018). While classical theories propose the existence of a built-in module that senses numerosity independently from other quantities (Dehaene, 2011), others argue that human quantification is rooted in a generalized magnitude system that treats various sensory streams indiscriminately to compute holistic representations (Leibovich et al., 2017). And finally, it remains unclear whether pitch chroma describes a biologically-determined perceptual property or rather a high-order construct of Western culture (Jacoby et al., 2019).

Although concerning different domains, these three debates are connected by a common thread. Namely, all scholars conceive the existence of "low-level" features that are presumably straightforward to process, at any age. Variables such as the envelope characteristics of a speech sound, the brightness of a visual array, the loudness of a tone or its fundamental frequency are generally labelled as "low-level". Meanwhile other variables are disregarded from being features of this kind even when they appear to be minimal components of adult perception. This is the case, among others, of phonetic units, approximate numerosity and pitch chroma which are (often tacitly) assumed to be *algorithmic* (Marr, 1982) percepts.
Such a conception arises by the simple observation that phonemes, numerosities and octaves are not physical entities, retrievable from the input in a "pre-packaged" way. That is to say, whereas to represent low-level features a direct mapping of a sensory variable seems sufficient, representing phonetic units, numerosity and pitch chroma might require a sequence of more complex operations.

Because inexperienced and physically immature, the infant brain in the first weeks after birth appears unequipped to perform the (presumably) necessary chain of transformations.
To start with, phonetic perception has been proposed to rely on articulatory knowledge (e.g. motor models), deriving from the development of production skills, that might be systematically confronted with the incoming speech signal (Kuhl et al., 2014). Switching to the topic of pitch quality, the perception of octave equivalence could result from acculturation (Frances & Dowling, 2014): Western grown-ups might have learned, consciously or unconsciously, a musical grammar where tones standing octaves apart are used interchangeably. Within these propositions, phonetic features and pitch chroma are "algorithmic" in that they implicate a chain of complex processes: acquisition, storage and comparison. It is natural to assume that very young brains would rather rely on more barren percepts (e.g. sound envelope, formant frequencies), which will be *enriched* thanks to exposure and maturation as they grow older[20]. Also numerosity might be conceived as an inferential construct, although with a slightly different connotation. As for the other elements, numbers do not exist in the external world in the way we, as educated adults, think of them. The estimation, even approximate, of numerosity (i.e. discrete quantity) might require the integration of multiple continuous dimensions that are not numerical per se (Gebuis et al., 2016). If that was the case, a refined process would be involved, whereby various sensory cues are weighted on the basis of their relevance and interrelation. Once again, a complex procedure seems required, making numerosity "algorithmic" and thus not viable for an immature and inexperienced brain. Speaking of, according to several scholars, infant quantitative behavior might reflect simple reactions to intensity or might rely on a rough summation of all sensory cues that can ultimately enable relational and generalized "more/less" judgments but nothing more precise (e.g. Mix et al., 2002, 2016; Leibovich et al., 2017). Only experience will lead the child to understand the relations between the various quantitative dimensions characterizing the physical world, enabling their *differentiation* and, ultimately, the formation of adult-like numerical concepts.

Yet, what appears algorithmic to the mind of highly educated adults (e.g. scientists and theoreticians who aim at understanding mental functions[21]) might not be for our perceptual systems. To explicate how this is the case let us take as an example time-to-contact. Similarly to phonetic perception and

[20] This family of ideas finds its roots in classical constructivist conceptions. For instance, according to Helmholtz the world we perceive comes about an act of imagination, associating current sensations to the ones from memory. He conceived perception as a process of unconscious inference (from sensory data), with a crucial prerequisite: abstract perceptual rules need to be acquired, inductively, from experience.

[21] Curious and pertinent to the present discussion the observation that not only naïve American adults, Indian adults and 6/7-year-olds but also psychologists and neuroscientists assume core neonatal skills to require years and learning to emerge. Among others, "thinking that an array of 10 items is more numerous than an array of 5" is generally considered as an acquired ability appearing in between 2 and 4 years of age (Wang & Feigenson, 2019). History of science teaches how folk beliefs can be hard to individuate or dismantle, such that they can permeate scientific theorizing of professionals even when they are clearly wrong.

numerical approximations, encoding time-to-contact is essential in every-day life, e.g. a crucial piece of information when an object is looming or when the observer approaches a surface/object that must be reached or avoided. Yet, getting to know when an object will contact the observer appears computationally complex, requiring a chain of calculations. To start with, the system needs to estimate the distance of the object; to do so, the real size of the object must be retrieved from memory and compared to the current size of the object projected onto the retina. Such an operation must be repeated after a known time interval, in order to estimate velocity. Finally, if velocity is constant, time-to-contact can be derived from the integration of distance and velocity. According to such implementation account, this percept is algorithmic in that it requires experience and memory, timekeeping and operations on intermediate representational codes (for e.g. visual angles and distances). Instead, the visual system seems to rely on the ratio between the optical position of the object's boundary and its rate of expansion (Lee, 1976; Lee & Reddish, 1981). The latter corresponds to a mathematical constant that specifies time-to-contact in a single step, without the need of computational chains involving retrieval from memory or intermediate representations of size or distance (Lee, 1974). The take-home message of this example is that our brain might be wired to capture "higher-order" (e.g. relational) variables directly, by means of detector mechanisms that implicate a single representation: their output. At least concerning approximate numerical estimations, the most recent neuroimaging investigations carried on adults provide evidence precisely in this direction: the adult visual cortex processes numerosity as a primary feature, without relying on intermediate representations of other magnitudes (e.g. Van Rinsveld et al., 2020, 2021). Crucially, since mechanisms of this sort do not require experience or higher-order computations that only a mature neural architecture might afford, there is not obvious reason to assume their unavailability early in life.

The existence of higher-order feature detectors might biologically grounded in (the principles underlying) population coding. As mentioned in the introduction, thanks to the advancement of multi-electrode arrays and optical/voltage imaging techniques, the last two decades of research have demonstrated that in many brain regions the essential unit of computation are neural populations (Saxena & Cunningham, 2019). The representational power of a neuronal ensemble is maximized by the possibility to combine various selectivity profiles, of e.g. single neurons, within a sophisticated and yet unitary code. Thus, within the research projects presented in this thesis, the capability of multivariate decoders to capture the content of macro-population codes have been particularly precious.

5.2. Our results in a nutshell

By exposing 3-month-olds human subjects to carefully calibrated multidimensional spaces composed of natural stimuli, by recording their neural responses with high-density EEG and by analyzing the collected data with pattern-analysis techniques, we have been able to provide novel

insights to solve all of the three debates discussed in the previous section. Concerning the first, we have uncovered how the brain, already in its early infancy, breaks down the speech input into orthogonal axes akin to the phonetic features described by linguists. Such a neural strategy creates a structured and compositional space robust to surface variability across voice peculiarities and co-articulatory contexts, not affordable by a plain domain-general spectrotemporal analysis of the input (Chapter 2). Shedding light upon the second debate, we have demonstrated that young infants encode numerical quantity spontaneously and clearly beyond the other concurrent quantitative characteristics of the input. Charmingly, whereas its abstractness was only inferable from adult psychophysics before, we could document the existence of a code for approximate numerosity that transcends sensory modality, presentation format (temporal/spatial) and arousal state (Chapter 3). Finally, we assessed that the infant brain characterizes pitch along its two psychological dimensions, height and chroma, and that within a predictable stimulus context (repeating tones) only the second keeps being tracked throughout time. Such observations clarify how pitch chroma corresponds to a basic organizing principle of neural responsivity rather than a product of culture (Chapter 4).

Altogether, what our results demonstrate is extraordinary simple: among the fundamental units used by young infants to encode and organize the sensory input there are not only "low-level" parameters, such as rate, duration and fundamental frequency but also "higher-order" dimensions. That is to say, **phonetic features, approximate numerosity and pitch chroma hold the role of representational primitives for the human brain**. Here the term *primitive* depicts the fact that these units are extracted automatically, i.e. without attentional engagement, and obligatorily, i.e. without any contingent necessity (e.g. externally induced needs such as a behavioral reaction). The term *primitive* further alludes to the early onset of such extraction mechanisms, observed here at ~12 weeks of age.

5.3. A common line between our results

Despite pertaining to distinct facets of our environment, these three primitives are very similar in respect to three intimately interconnected key aspects.

a. Phonetic features, numerosity and chroma all imply **dimensionality reduction**. Any brain faces the fundamental challenge of processing high-dimensional stimuli with a set of biological resources that is unavoidably limited by metabolic cost (Lennie, 2003) and anatomical bottlenecks (just to have an idea: the input of 100 million photoreceptors must be passed to 1 million optic nerve fibers). One way to address this challenge[22] is to describe a particular stimulus space by means of a restrained set of variables, thereby compressing its dimensionality. Crucially,

[22] The reader may notice that structural limitations might be even more challenging early in development due to structural immaturity (de Graaf-Peters & Hadders-Algra, 2006)

all the representational primitives under discussion enable low-dimensional embedding of high-dimensional sensory data. Specifically, a phonetic code can describe an intricate spectrotemporal pattern with a handful of parameters. Numerosity summarizes multiple streams of continuous information into a single value. Chroma captures the common quality of many fundamental frequencies (our hearing apparatus spans 10 octaves!).

b. As mentioned in the incipit of this thesis, complexity does not reside solely in dimensionality but also, and even to a greater extent, in the extreme variability of the outside world. The neural codes isolated by our experiments appear particularly useful in this regard, as they capture **invariance**. By definition, phonetic features and phonemes remain always the same, irrespective of the ongoing acoustic context (speaker's voice, intonation, co-articulation). The same approximate numerosity can take innumerous material forms depending on the sensory modality at hand and the particular physical variables proper of each modality (e.g. the size and the spatial disposition of visual items or the duration and rate of auditory events). Lastly, complex tones will change considerably in their spectral composition depending on their emitter, yet they may share the same quality. Precisely due to such phenomenon, it is possible to identify similar prosodies across speakers and the same melodies when played by distinct instruments.

c. The phonetic, numerical and chromatic codes are similar in a third aspect: they clutch information that is **highly relevant** for our species. Phonetic features correspond to the smallest units carrying linguistic meaning (Stevens, 2002). An approximate sense of numerosity is likely to be the cradle of arithmetic competence (Butterworth, 2018; Odic & Starr, 2018), the latter permeating innumerous aspects of our life since childhood. And chromatic patterns covey intonation (e.g. distinguish questions from exclamations and regulate conversational interaction) and emotions thereby holding an inestimable communicative value. Overall, these discrete codes enable to reach representational abstractions that are maximally adaptive for the human being.

In summary, our results demonstrate how, despite what could be intuitively foreseen by scholars, early encoding mechanisms go beyond a plain mapping of physical properties. Rather, the infant brain spontaneously redescribes incoming information into strategic formats that, by reducing dimensionality and overcoming variability, provide structure. This process allows characterizing relevant facets of the external world by means of encoding primitives that afford extreme representational **flexibility** and are therefore advantageous for further processing. For example, the breath of the human lexicon relies precisely on the combinatorial possibilities of phonetic features (aka a small set of units); gathered into phonemes, and words thereafter. Switching to quantities, non-numerical descriptors are confined to only a subset of sensory modalities; in contrast, numerosity can afford representations pertaining to any sensory modality. Moreover, unlike e.g. rate (temporal) or e.g. density (spatial), a numerical sum can pool information distributed over both time and space.

5.4. An ideal ground for learning

One of the fundamental goals of cognitive science is to explain how humans acquire knowledge. When trying to delineate a framework of infant learning, one of the major challenges encountered consists in elucidating how young brains confront "combinatorial explosions" (Malsburg, 1995). Namely, an efficient mechanism of knowledge acquisition must be able to deal with the fact that the amount of potentially relevant information increases exponentially with the number of features in the data. The latter accumulates rapidly as the sensory environment faces continuous evolution. Despite this burdensome computational problem infants are formidable learners. How is that possible?

One explanation could reside in the existence of a set of implicit constraints determining the acquisition of minimally sufficient rather than complete representations of the input (Aslin & Fiser, 2005). For example, it has been pointed out that developmental factors might be at play: limited working memory and sensory immaturity can effectively reduce the number of elements that needs to be taken into account while attempting to discern rules and regularities (e.g. Newport, 1990; Elman, 1993) . We believe that the peculiar encoding strategies isolated by our experiments can be conceived, in a way, as being part of these constraints, functioning as catalyzers for learning. As a matter of fact, the representational processes uncovered by our studies allow to minimize the complexity of the world by creating discrete and circumscribed spaces composed of ecologically relevant variables that can be further manipulated and combined with ease.

In support of our proposition, the experimental data collected so far testifies how young infants can indeed use the encoding primitives under discussion to learn structure. As a reference let us take distributional learning (DL), a form of statistical learning that entails the acquisition of knowledge based on the distributional structure of the encountered input. In DL experiments, subjects are exposed to stimuli that vary in equal steps along a particular dimension, thereby forming a continuum. The stimuli are presented with frequencies that constitute either a bimodal or a unimodal distribution, in which the tokens near the endpoints or around the middle of the continuum (respectively) are most frequent. Maye et al. (2002) used this paradigm with 6- and 8-mo olds, exposing them to eight tokens forming a continuum from "da" to "ta" which varied in equal steps along the dimension of voice-onset-time (VOT). They observed that only those infants who listened to the bimodal distribution could reliably discriminate its endpoints in a subsequent test phase. Crucially, in a follow-up experiment, Maye, Weiss and Aslin (2008) showed that after exposure to a bimodal distribution from "da" to "ta" (characterized by the place of articulation "alveolar") infants could successfully discriminate the VOT contrast under study not only when tested on the same class of stimuli but also within an alternative phonetic context ("ka" vs "ga", defined by the place of articulation "velar"). These observations demonstrate that learning pertained to the phonetic-feature level, which is possible only when the incoming speech input is encoded phonetically. Importantly, computational modelling reveals that what subjects acquire from frequency distributions are not the phonetic distinctions themselves (e.g. VOT). Rather, they learn to group together isolated regions of

pre-existing dimensions into unified categories (McMurray et al., 2009). This implicates that phonetic encoding is a prerequisite for and not the end product of the acquisition process. Distributional learning is thought to be one of the key mechanisms enabling infants do discover the phonetic repertoire of their native tongue (Werker et al., 2012).
Similarly to what observed for the sensitivity to phonetic contrasts (Werker & Tees, 1984), the initial ability of infants to distinguish a broad variety of lexical tones is reshaped by ambient language over the course of the first year (e.g. nonnative tone discriminability tends to decline; Mattock et al., 2008). Intriguingly, the same DL process operating at the phonetic level might enable tone-language learners to individuate those specific pitch variations defining the native repertoire of lexical tones. A first hint in this direction has been brought by Liu and Kager (2014), who used the paradigm described above on 11/12-mo-olds. They showed that exposition to a bimodal continuum could re-instantiate the discriminability of a tonal contrast no longer detectable after the age of 7-8 months. Notably, for this learning process to succeed in realistic settings, lexical tones must be processed in terms of F0 quality rather than F0 absolute value since only the former code is robust to changes in timbre (i.e. speakers). Finally, Libertus and colleagues (2018) have recently reported that 6-mo-olds familiarized with a wide range of dot arrays whose numerosity was centered around either a single mean or two peaks spontaneously extracted the distributional structure of the input received.

While the studies reviewed in this section demonstrate how infants are capable of employing phonetic and numerical codes to acquire new knowledge, our projects show that these codes are engaged in a completely automatic fashion at a much younger age and when there is no structure to learn (i.e. when naturally rich stimuli are presented in a randomized order). The two lines of research complement each other: whereas we have revealed the neural reality of these codes, several behavioral studies have demonstrated how these codes can be used as a basis to learn important properties of the environment within the first year of life.

5.5. Innate?

Although babies were once thought to be blank slates, infinitely malleable, the last decades of research has brought dozen of findings illustrating how such belief does not hold realistic. Given these demonstrations, many have invoked the concept of innateness referring, among others, to a *language instinct* (e.g. Pinker, 1994) and a *sense for number* (e.g. Dehaene, 2011). Are the neural encoding mechanisms isolated by our experiments innate?
Any friendly reviewer would certainly notice that whereas "innate" is often meant as "present at birth" we did not test newborns. At this point, we may argue that our evidence is corroborated by the spontaneous tendency of neonates to match visual displays and auditory sequences containing an equal number of items (Izard et al., 2009; Coubart et al., 2014), indicating that the abstract code for discrete quantity isolated through our paradigm is likely to be available and operative since start.

Still, presence at birth is neither necessary[23] nor sufficient when the term "innate" is interpreted as "not learned". Embracing such a perspective, the friendly reviewer would highlight that learning starts before birth, speaking with full knowledge of the facts (James, 2010). Now, it would be meaningful to assert that phonetic discrimination has been documented in neonates born at a gestational age of 30 weeks. In these preterms, the detection of a phonetic change is supported by a network of brain areas distinct from the regions carrying speaker-related processing and resemblant the functional pathways recruited later on (Mahmoudzadeh et al., 2013). Such observations corroborate our findings on pre-babblers in depicting genuine phonetic encoding as a genetically determined neural equipment.

Actually, the friendly reviewer is quite precise, the cochlea is anatomically functional at 20 weeks of gestation (Graven & Browne, 2008) and the maturation of the brain stem auditory pathway is quite advanced at this stage of fetal life, featuring fully-grown (although non-myelinated) axons and stainable dendritic arbors (Eggermont & Moore, 2012). There might be a self-organizing mechanism that from the 20th week of pregnancy shapes the neural circuitry in course of formation according to the characteristics of the external auditory environment (Bharucha & Mencl, 1996). Since for a human mom the latter is largely composed of harmonic sounds, i.e. speech and music, this mechanism could account for the *acquisition* of octave categories, thereby explaining the ubiquity of octave equivalence across cultures. Those who assumed that language and music were creations of the human kind, presenting a harmonic structure due to the perceptual predispositions of our own species, now might wonder who created language and music for us.

Since the innateness concept is irretrievably confused (for enlightening commentary on this topic see Griffiths, 2002 and Samuels, 2004), other scientific domains have come to eschew it altogether. Yet, in cognitive science whether a trait is innate is still regarded as a significant question. Instead of entering vicious cycles as the one just outlined, we should perhaps consider the reason why this is the case. The importance attributed to innateness by cognitive scientists concerns the theoretical commitments of the discipline itself. Under pain of regress, any cognitive theory must presuppose the availability of a set of structures and resources. Once posited, the latter can be invoked by the theory to elaborate the emergence of new ones (Samuels 2004). Cognitive development is necessarily grounded in the disposal of abilities for detecting and analyzing inputs and for drawing inferences. In extreme empiricist theories, the initial state of such abilities is held to consist of sensory transducers and one or more domain-general learning mechanisms. Yet, this type of characterizations provide no convincing account of either the rapidity of early learning or the lack of massive interference between the various conceptual domains over which development operates (Spelke & Kinzler, 2009). As discussed in the previous paragraphs, neural codes for phonetic units, numerosity and pitch chroma afford to represent meaningful aspects of the world in a discrete and

[23]As a matter of fact, a given capacity can emerge over the course of development due to maturation rather than experience.

abstract format. Given such key characteristics, these encoding mechanisms can provide a more satisfactory account for the rapid pace and flexibility of early development relatively not only to radical constructionists views but also to previous proposal pertaining to holistic syllables or broad spectrotemporal patterns, continuous magnitudes or aspecific intensity estimations and disorganized tonal frequency components.

To our knowledge, there is currently no cognitive process that can adequately explain their appearance, such that phonetic features, numerosity and pitch chroma may be considered *primitives* also in the psychological sense (Samuels, 2002)[24]. That is, the explanation for their acquisition[25] may be found at some other level, be it molecular or neurobiological. Still, the latter is just a "bonus" consideration. Most importantly, even if the appearance of these encoding mechanisms was explicable by means of some neurocognitive phenomenon, we propose that their early onset and full automaticity/obligatoriness are striking enough for them to be regarded as foundational building-blocks of human cognition, without the need to invoke innateness. In fact, their delineation provides a new piece of understanding to the puzzle of the origins of knowledge in general (as summarized above) and novel inspiration to foster future research programs.

5.6. Future directions

Beyond providing deeper insights upon the origins of human cognition, the present thesis sets a promising methodological ground for new projects in the field of developmental neuroscience.

The methodological value and innovation of our paradigms resides in the *combination* of two key elements: **multidimensional stimuli spaces** and strategical multivariate **cross-condition decoding.**

Since relevant features have been manipulated singularly, in most of the previous investigations on infant representational skills each neural response could be analyzed in respect of a single variable of interest. These paradigms enabled to reveal whether the infant brain is capable of detecting the regularity of a certain characteristic and whether changes pertaining to distinct dimensions are associated to mismatch responses over distinct brain areas. This type of paradigm would have been incompatible with our goal of capturing the encoding mechanisms engaged in ecological settings. To fulfill our purpose, we constructed multidimensional spaces where the natural co-occurrence or interrelation between various features could be maintained. Since in this experimental setting each

[24] According to Samuels, "a concept, belief, learning mechanism or module is a psychological primitive when there is no correct scientific psychological theory that explains its acquisition".

[25] Here we refer to « acquisition » in its baseline sense: a structure S is acquired by an object O if and only if O fails to possess S at all times prior to time t, but possess S at t. When the idea of acquisition is considered in this minimal sense, all cognitive structures must be labelled as acquired.

sub-condition is definable along more than one dimension, the mere possibility to classify stimulus identity from neural responses provides limited understanding of the underlying encoding scheme. Instead, through our analyses, we have shown how cross-condition decoding can be used as a powerful technique to uncover the nature of the neural codes, provided that the experimental stimulation itself is carefully designed in accordance to this goal. Within our first project, the pattern of generalization enabled us to discern whether phonetic dimensions are encoded orthogonally or rather conjointly, thereby revealing the availability of a structured, compositional code for speech (Chapter 2). Within our second project we have shown how cross-decoding enables to overcome physical obstacles: capitalizing on a strategic combination of trainings and tests allowed us to overcome the impossibility of presenting different numbers of items without non-numerical covariations. Finally, while through cross-condition analyses fMRI adult studies could isolate multi-modal object representations (Man et al., 2015) and consistent neural activity between imagery and perception (Cichy et al., 2012), we individuated an abstract code for numerosity that generalizes across sensory modalities *and* formats (from temporal to spatial displays).

To recapitulate, the use of high-density EEG and multivariate cross-classification combined to rich, carefully calibrated stimulus spaces can provide crucial insights upon the encoding strategies of the brain. Currently in our lab, we are employing this methodological framework in order to broaden our characterization of early speech processing. In our follow-up experiment, we present a new set of syllables (Figure 5.1), once again in a randomized (Latin squared) order, to neonates while their neural responses are recorded with a 128-channel EEG system. The goal of this study is to extend the findings described in Chapter 2 along multiple lines. First, we will try to replicate the possibility to decode consonantal phonetic features (VOT and place of articulation; Figure 5.1 top left) in a younger population and from neural data with a reduced spatial resolution (128 EEG channels instead of 256). Second, we will investigate the encoding of vocalic phonetic features (height and backness; Figure 5.1 top right), the occurrence of which could be only inferred in our previous study. Crucially, we will use cross-condition analyses to ask whether featural and/or phoneme encoding are independent from the position of consonants and vowels within the syllable (Figure 5.1; bottom).

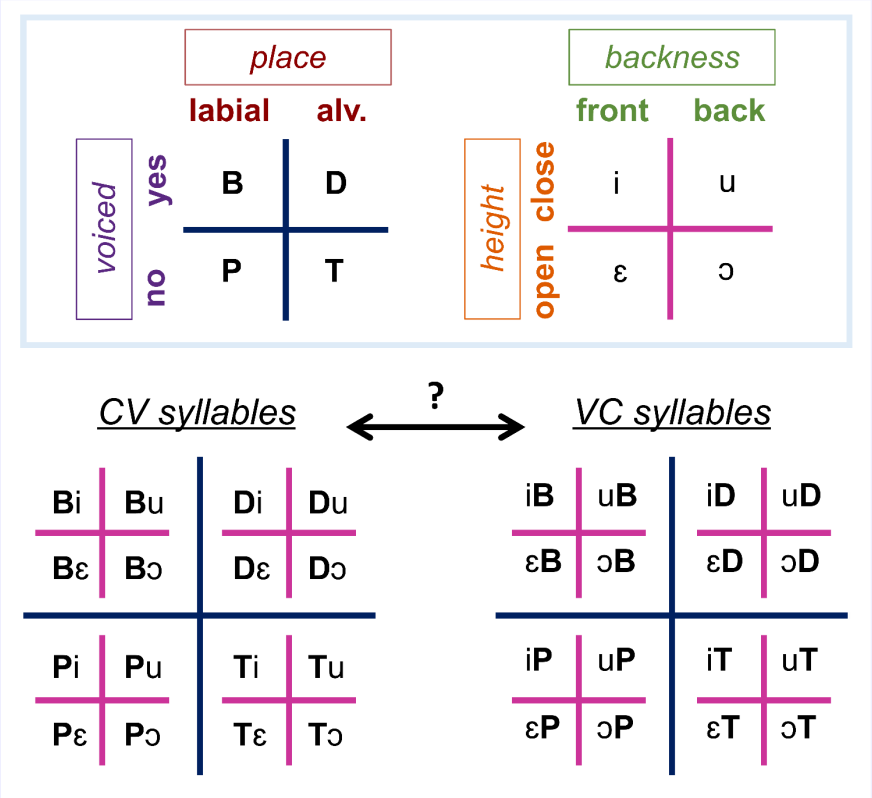

Figure 5.1 Current experimental paradigm to investigate speech processing in neonates. To ensure natural variability, syllables are spoken by four alternative female voices from different nationalities for a total of 128 distinct tokens. Alv.=alveolar, "voicing" refers to voice-onset-time (VOT; yes=voiced, no=voiceless).

More broadly, we believe that our methodological approach will be ideally suited to delineate representational primitives in other perceptual domains. For instance, according to Spelke, the origins of human knowledge reside in a small set of foundational cognitive systems, each centered on a few key principles that serve to individuate the entities in its domain and draw inferences about them (Spelke & Kinzler, 2007). Drawing a connection with the way we interpret our own empirical observations, the systems proposed under the core knowledge view may function as catalyzers for human cognitive development because they are based on principles that are veridical and adaptive at the scale at which humans perceive and behave. Within this framework, our ability of identifying token objects or objects of particular kinds might originate from a core system of geometrical intuitions pertaining to small-scale visual configurations (Spelke, 2011). In this regard, a recently collected body of data (Izard et al., 2021) shows that, in absence particular instructions, preschoolers and adults from both the U.S. or the Amazon analyze 2D visual forms based on Euclidean metric properties, i.e. in terms of shape (defined by angles and length proportions) and global size, disregarding changes in position, orientation and reflection. Considering evidence of this sort, it has been highlighted that shape and size might constitute representational pillars for universal

geometrical intuitions (Izard et al., 2021). The methodological approach developed by the present thesis would enable to investigate the existence of such representational pillars early in infancy. To start with, whereas behavioral evidence indicates that not only infants but even children have limited ability to perceive shape-defining angles (Dillon et al., 2020; Izard et al., 2011, 2014), it would be possible to ascertain whether such findings reflect a true computational limit or rather a lack of methodological sensitivity (as it was the case for e.g. the apparent lack of sub-syllabic speech encoding on the behalf of pre-babblers). Mainly, the employment of a multidimensional visual space combined with cross-condition decoding would be ideally suited to ask whether young infants process shape- and/or size-related information irrespective of position, orientation and reflection (i.e. according to Euclidean principles). Further, it would be possible to elucidate the elementary units of encoding: global size might be computed holistically, e.g. in terms of implied area, or rather compositionally, through minimal codes for line lengths and distance between extremities. Overall, the methodology we propose would enable to query the existence and characteristics of representational primitives akin to abstract geometrical elements.

Investigation within new domains might implicate a more extensive use of visual stimulation and thus an additional challenge: infants will need to be tested awake (obviously) and they will need to direct their sight toward the stimulation screen, which requires a quite high degree of collaboration. Crucial aid in this setting could be brought by the employment of the rapid serial visual presentation (RSVP) paradigm, where stimuli are flashed in very fast sequences (typically 10/s in adult studies). With such a rapid presentation, picture recognition remains possible for adults and selective neural responses are preserved in the macaque temporal cortex (Keysers et al., 2001), although subjective visibility is degraded and often stimuli remain unperceived at the conscious level (Lawrence, 1971). This strategy is very promising for the study of representational primitives within the visual modality because enables to collect a big amount of trials in a short period[26]. Moreover, it grants the investigation of automatic processing, as it allows no time for higher-order attentional phenomena to occur. Our research group, including the writer, have started to work on this strategy to investigate the perception of visual categories (faces, bodies, objects and houses) with MEG in 14-week-olds. Unfortunately, the project has been interrupted by the COVID-19 pandemic when pilot sessions were starting to yield encouraging outcomes.

Finally yet importantly, the paradigms proposed here might open new opportunities for the clinical field. As mentioned in Chapter 1, prominent inter-subject variability makes ERP components unreliable when it comes to determine whether an individual is developing abnormally (Picton & Taylor, 2007). Because they are carried within subject, multivariate pattern analyses offer an alternative option: the possibility to collect normative data concerning psychological, rather than

[26] The "collaboration" of a young infant will not last more than ~10 minutes while multivariate estimators require a high number of samples to reach stable decision boundaries.

physiological, variables in order to detect dysfunction. Concretely, once a given encoding primitive (including its level of abstraction and temporal dynamics) has been delineated, it is possible to investigate potential alterations related to developmental disorders. For instance, the implementation of our paradigms[27] within longitudinal studies on subjects at risk of dyslexia, dyscalculia or amusia might bring useful insights and, ultimately, to the definition of effective criteria for early diagnosis.

[27] An additional advantage of our paradigms is the possibility to detect phonetic and neural codes during sleep, i.e. with minimal collaboration required on the behalf of the subject. However, this benefit is restricted to paradigms relying on auditory stimuli.

5.7. Bibliography

Aslin, R., & Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *Trends in Cognitive Sciences*, *9*(3), 92–98. https://doi.org/10.1016/j.tics.2005.01.003

Bharucha, J. J., & Mencl, W. E. (1996). Two Issues in Auditory Cognition: Self-Organization of Octave Categories and Pitch-Invariant Pattern Recognition. *Psychological Science*, *7*(3), 142–149. https://doi.org/10.1111/j.1467-9280.1996.tb00347.x

Butterworth, B. (2018). The implications for education of an innate numerosity-processing mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170118. https://doi.org/10.1098/rstb.2017.0118

Cichy, R. M., Heinzle, J., & Haynes, J.-D. (2012). Imagery and Perception Share Cortical Representations of Content and Location. *Cerebral Cortex*, *22*(2), 372–380. https://doi.org/10.1093/cercor/bhr106

Coubart, A., Izard, V., Spelke, E. S., Marie, J., & Streri, A. (2014). Dissociation between small and large numerosities in newborn infants. *Developmental Science*, *17*(1), 11–22. https://doi.org/10.1111/desc.12108

de Graaf-Peters, V. B., & Hadders-Algra, M. (2006). Ontogeny of the human central nervous system: What is happening when? *Early Human Development*, *82*(4), 257–266. https://doi.org/10.1016/j.earlhumdev.2005.10.013

Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition*. Oxford University Press, USA.

Dehaene-Lambertz, G. (2017). The human infant brain: A neural architecture able to learn language. *Psychonomic Bulletin & Review*, *24*(1), 48–55. https://doi.org/10.3758/s13423-016-1156-9

Dillon, M. R., Izard, V., & Spelke, E. S. (2020). Infants' sensitivity to shape changes in 2D visual forms. *Infancy*, *25*(5), 618–639. https://doi.org/10.1111/infa.12343

Eggermont, J. J., & Moore, J. K. (2012). Morphological and Functional Development of the Auditory Nervous System. In L. Werner, R. R. Fay, & A. N. Popper (Eds.), *Human Auditory Development* (pp. 61–105). Springer. https://doi.org/10.1007/978-1-4614-1421-6_3

Elman, J. L. (1993). *Learning and development in neural networks: The importance of starting small*.

Frances, R., & Dowling, W. J. (2014). *The Perception of Music*. Psychology Press.

Frauenfelder, U. H., & Floccia, C. (1998). The Recognition of Spoken Word. In A. D. Friederici (Ed.), *Language Comprehension: A Biological Perspective* (pp. 1–40). Springer. https://doi.org/10.1007/978-3-642-97734-3_1

Gebuis, T., Cohen Kadosh, R., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta Psychologica*, *171*, 17–35. https://doi.org/10.1016/j.actpsy.2016.09.003

Graven, S. N., & Browne, J. V. (2008). Auditory Development in the Fetus and Infant. *Newborn and Infant Nursing Reviews*, *8*(4), 187–193. https://doi.org/10.1053/j.nainr.2008.10.010

Griffiths, P. E. (2002). What Is Innateness? *The Monist*, *85*(1), 70–85. https://doi.org/10.5840/monist20028518

Izard, V., O'Donnell, E., & Spelke, E. S. (2014). Reading Angles in Maps. *Child Development*, *85*(1), 237–249. https://doi.org/10.1111/cdev.12114

Izard, V., Pica, P., & Spelke, E. (2021). *Visual Foundations of Euclidean Geometry*. PsyArXiv. https://doi.org/10.31234/osf.io/rmdeh

Izard, V., Pica, P., Spelke, E. S., & Dehaene, S. (2011). Flexible intuitions of Euclidean geometry in an Amazonian indigene group. *Proceedings of the National Academy of Sciences*, *108*(24), 9782–9787. https://doi.org/10.1073/pnas.1016686108

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, *106*(25), 10382–10385. https://doi.org/10.1073/pnas.0812142106

Jacoby, N., Undurraga, E. A., McPherson, M. J., Valdés, J., Ossandón, T., & McDermott, J. H. (2019). Universal and Non-universal Features of Musical Pitch Perception Revealed by Singing. *Current Biology*, *29*(19), 3229-3243.e12. https://doi.org/10.1016/j.cub.2019.08.020

James, D. K. (2010). Fetal learning: A critical review. *Infant and Child Development*, *19*(1), 45–54. https://doi.org/10.1002/icd.653

Kazanina, N., Bowers, J. S., & Idsardi, W. (2018). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*, *25*(2), 560–585. https://doi.org/10.3758/s13423-017-1362-0

Keysers, C., Xiao, D.-K., Földiák, P., & Perrett, D. I. (2001). The Speed of Sight. *Journal of Cognitive Neuroscience*, *13*(1), 90–101. https://doi.org/10.1162/089892901564199

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J.-F. L., & Imada, T. (2014). Infants' brain responses to speech suggest Analysis by Synthesis. *Proceedings of the National Academy of Sciences*, *111*(31), 11238–11245. https://doi.org/10.1073/pnas.1410963111

Lawrence, D. H. (1971). Two studies of visual search for word targets with controlled rates of presentation*. *Perception & Psychophysics*, *10*(2), 85–89. https://doi.org/10.3758/BF03214320

Lee, D. N. (1974). Visual information during locomotion. In *Perception: Essays in honor of James J. Gibson* (pp. 317–317). Cornell University Press.

Lee, D. N. (1976). A Theory of Visual Control of Braking Based on Information about Time-to-Collision. *Perception*, *5*(4), 437–459. https://doi.org/10.1068/p050437

Lee, D. N., & Reddish, P. E. (1981). Plummeting gannets: A paradigm of ecological optics. *Nature*, *293*(5830), 293–294. https://doi.org/10.1038/293293a0

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16000960

Lennie, P. (2003). The Cost of Cortical Computation. *Current Biology*, *13*(6), 493–497. https://doi.org/10.1016/S0960-9822(03)00135-0

Libertus, M. E., Feigenson, L., & Halberda, J. (2018). Infants Extract Frequency Distributions from Variable Approximate Numerical Information. *Infancy*, *23*(1), 29–44. https://doi.org/10.1111/infa.12198

Liu, L., & Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition*, *133*(2), 385–394. https://doi.org/10.1016/j.cognition.2014.06.004

Mahmoudzadeh, M., Dehaene-Lambertz, G., Fournier, M., Kongolo, G., Goudjil, S., Dubois, J., Grebe, R., & Wallois, F. (2013). Syllabic discrimination in premature human infants prior to complete formation of cortical layers. *Proceedings of the National Academy of Sciences*, *110*(12), 4846–4851. https://doi.org/10.1073/pnas.1212220110

Malsburg, C. von der. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, *5*(4), 520–526. https://doi.org/10.1016/0959-4388(95)80014-X

Man, K., Damasio, A., Meyer, K., & Kaplan, J. T. (2015). Convergent and invariant object representations for sight, sound, and touch: Neural Convergence of Sight, Sound, and Touch. *Human Brain Mapping*, *36*(9), 3629–3640. https://doi.org/10.1002/hbm.22867

Marr, D. (1982). *Vision*. Freeman.

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106*(3), 1367–1381. https://doi.org/10.1016/j.cognition.2007.07.002

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. https://doi.org/10.1016/S0010-0277(01)00157-3

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. https://doi.org/10.1111/j.1467-7687.2009.00822.x

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). Multiple cues for quantification in infancy: Is number one of them? *Psychological Bulletin*, *128*(2), 278–294. https://doi.org/10.1037/0033-2909.128.2.278

Mix, K. S., Levine, S. C., & Newcombe, N. S. (2016). Chapter 1—Development of Quantitative Thinking Across Correlated Dimensions. In A. Henik (Ed.), *Continuous Issues in Numerical Cognition* (pp. 1–33). Academic Press. https://doi.org/10.1016/B978-0-12-801637-4.00001-9

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28. https://doi.org/10.1016/0364-0213(90)90024-Q

Odic, D., & Starr, A. (2018). An Introduction to the Approximate Number System. *Child Development Perspectives*, *12*(4), 223–229. https://doi.org/10.1111/cdep.12288

Picton, T. W., & Taylor, M. J. (2007). Electrophysiological Evaluation of Human Brain Development. *Developmental Neuropsychology*, *31*(3), 249–278. https://doi.org/10.1080/87565640701228732

Pinker, S. (1994). *The language instinct*. William Morrow & Co.

Samuels, R. (2002). Nativism in Cognitive Science. *Mind & Language*, *17*(3), 233–265. https://doi.org/10.1111/1468-0017.00197

Samuels, R. (2004). Innateness in cognitive science. *Trends in Cognitive Sciences*, *8*(3), 136–141. https://doi.org/10.1016/j.tics.2004.01.010

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, *55*, 103–111. https://doi.org/10.1016/j.conb.2019.02.002

Spelke, E. S. (2011). Natural Number and Natural Geometry. In *Space, Time and Number in the Brain* (pp. 287–317). Elsevier. https://doi.org/10.1016/B978-0-12-385948-8.00018-9

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Spelke, E. S., & Kinzler, K. D. (2009). Innateness, Learning, and Rationality. *Child Development Perspectives*, *3*(2), 96–98. https://doi.org/10.1111/j.1750-8606.2009.00085.x

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872–1891. https://doi.org/10.1121/1.1458026

Van Rinsveld, A., Guillaume, M., Kohler, P. J., Schiltz, C., Gevers, W., & Content, A. (2020). The neural signature of numerosity by separating numerical and continuous magnitude extraction in visual cortex with frequency-tagged EEG. *Proceedings of the National Academy of Sciences*, *117*(11), 5726–5732. https://doi.org/10.1073/pnas.1917849117

Van Rinsveld, A., Wens, V., Guillaume, M., Beuel, A., Gevers, W., De Tiège, X., & Content, A. (2021). Automatic Processing of Numerosity in Human Neocortex Evidenced by Occipital and Parietal Neuromagnetic Responses. *Cerebral Cortex Communications*, *2*(tgab028). https://doi.org/10.1093/texcom/tgab028

Wang, J. (Jenny), & Feigenson, L. (2019). Is Empiricism Innate? Preference for Nurture Over Nature in People's Beliefs About the Origins of Human Knowledge. *Open Mind*, *3*, 89–100. https://doi.org/10.1162/opmi_a_00028

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. https://doi.org/10.1016/S0163-6383(84)80022-3

Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How Do Infants Become Experts at Native-Speech Perception? *Current Directions in Psychological Science*, *21*(4), 221–226. https://doi.org/10.1177/0963721412449459