Orthogonal neural codes for speech in the infant brain

Giulia Gennari[a,1], Sébastien Marti[a], Marie Palu[a], Ana Fló[a], and Ghislaine Dehaene-Lambertz[a]

[a]Cognitive Neuroimaging Unit U992, Institut National de la Santé et de la Recherche Médicale, Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Direction de la Recherche Fondamentale/Institut Joliot, Centre National de la Recherche Scientifique *ERL9003,* NeuroSpin Center, Université Paris-Saclay, 91191 Gif-sur-Yvette, France

Edited by Richard N. Aslin, Haskins Laboratories, New Haven, CT, and approved June 25, 2021 (received for review October 7, 2020)

Creating invariant representations from an everchanging 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 electroencephalographic system, we recorded the neural responses of 3-mo-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.

speech | phoneme | infant | ERP | language

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 coarticulation. 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 (1) and face recognition (2), 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 (2). 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 (3). 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 versus +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 speech input, thereby overcoming its variability.

In the last years, high-resolution intracranial recordings on adults (4) and functional MRI (fMRI) adult data (5, 6) 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 (refs. 7 to 9, 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 (10, 11), speaking rate (12, 13), and prosody (14). 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 mo olds fail at identifying a shared consonant in a group of syllables containing different vowels (15, 16). Furthermore, neonates proved capable of categorizing utterances using the number of their syllabic

Significance

For adults to comprehend spoken language, and for infants to acquire their native tongue, it is fundamental to encode speech as a sequence of stable and invariant segments despite its extreme acoustic variability. We show that the brain of a 3-mo-old baby can achieve this critical task thanks to a decomposition system which breaks down the speech input into minimal and orthogonal components such as the manner and the place of articulation. These elementary units are robust to signal variability and are flexibly recombined into phoneme identities during a second processing phase. Our data indicate that a combinatorial neural code for speech is present at an early stage of language development.

Author contributions: G.G. and G.D.-L. designed research; G.G. and M.P. performed research; G.G. analyzed data; A.F. contributed new analytic tools; S.M. and G.D.-L. supervised data analysis; and G.G. and G.D.-L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: giulia.gennari1991@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2020410118/-/DCSupplemental.

Published July 29, 2021.

constituents but not the number of phonemes (17). Following these results, many authors have proposed the syllable as the primitive unit for speech processing. Computational modeling 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 (18). 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, electroencephalographic (EEG) recordings revealed that 3 mo olds could recognize the shared consonant and detect when it changed (19). They could even learn to associate each consonant to a visual shape independently of the vocalic surroundings (20). Such a finding, easily explicable in terms of sub-syllabic processing, prompts to reexamine the format of early speech representations.

To this aim, we combined high-resolution EEG recordings with time-resolved multivariate pattern analysis. A total of 25 3-mo-old infants were exposed to 120 natural CV syllables presented in pseudorandom order during about 1 h. Syllables were chosen to independently vary the consonantal dimensions of manner (obstruent versus sonorant) and place of articulation (labial versus alveolar versus 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 coarticulatory variability across tokens with the same phonetic profile (Fig. 1*A*). 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 spectro-temporal prototypes (21), the acoustic cues for place are more subtle (22) and complex (23), hence fundamentally dependent on the context of production (24). Such acoustical divergence was especially evident in the auditory similarity structure of our stimuli set, as illustrated in *SI Appendix*, Fig. S1.

We used multivariate decoding analyses to investigate infant speech processing at three possible levels corresponding to holistic syllables, phonemes, and phonetic features.* 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 (25). 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 nonpertinent for a given code (26). 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 subsyllabic level, we could ask whether a decoder trained to separate "bo" versus "do" is able to 1) correctly classify "mi" versus "ni," thus revealing the presence a neural code for the places "labial" versus "alveolar" that is orthogonal to vowels and manners, or 2) generalize only to "bi" versus "di," thus indicating an idiosyncratic and integrated neural code for the consonants "b" versus "d" without further decomposition into separable dimensions.

In addition, by 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 (27). By tracing the time course of generalizations and class confusability, we could ask whether and when particular pieces of information were recoded across stages of processing. A factorized encoding model, similar to that observed for faces (2), 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 ∼3,100 trials/participant. Furthermore, we collected ERPs with a high-density custom net featuring an unusual number of 256 channels (Figs. 1*B* and *SI Appendix*, Fig. S2; see also Fig. 1*C* and Movie S1 for the grand average across all syllables and its sources). This intensive electrode coverage, combined with the thinness of infant skulls, should enhance the spatial resolution of our recordings and facilitate the discrimination of ERPs arising from spatially close neuronal clusters (28).

Results

For all the analyses described below, we trained and tested series of linear estimators on brief (20 ms) 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?

Successful Classification Is Achieved on the Basis of 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. Fig. 2 *A* and *B* show that obstruents could be distinguished from sonorants starting from 80 ms after syllable onset ($p_{clust} = 0.0001$; peak performance observed at 200 ms: $n = 25$, $M = 0.735 \pm 0.08$, chance = 0.5), while places of articulation were reliably classified over two time windows: 220 to 480 ms ($p_{clust} = 0.0001$; peak at 260 ms: $M = 0.545 \pm 0.039$) and 540 to 720 ms ($p_{clust} = 0.0028$; peak at 640 ms: $M = 0.534 \pm 0.042$). 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 subclasses. Nonetheless, Fig. 2*C* shows that vowel identity was reliably discerned in between 260 and 600 ms ($p_{clust} = 0.0001$; peak at 480 ms: $M = 0.596 \pm 0.08$, chance = 0.5) and from 760 ms onwards ($p_{clust} = 0.0001$; peak at 860 ms: $M = 0.56 \pm 0.067$, 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 (27). Fig. 2*D* illustrates how classifiers generalized only over a limited amount of time lags, an 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. 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 (*SI Appendix, Weights Projection*). Discriminative activity was diffused over the scalp,

*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.

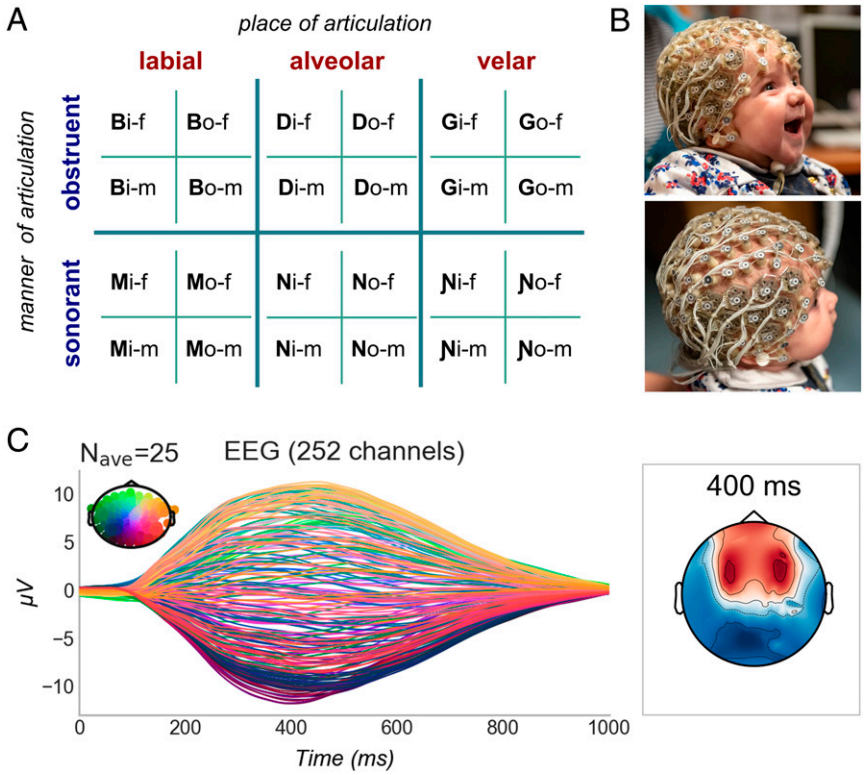

Fig. 1. Experimental setup and average syllable-related potential. (*A*) Stimuli subconditions and their phonetic characteristics (f = female, m = male voice). (*B*) 256 channels super high-density net on a 3-mo-old infant: tight grids of custom electrodes are arranged over the auditory linguistic areas of the superior temporal lobe (see also *SI Appendix*, Fig. S2). (*C*) Grand average ERP: all conditions are pulled together.

resembling the auditory ERP topographies arising from multiple perisylvian sources that are typical of this age (*SI Appendix*, Fig. S3 and Movie S1). Crucially, informative clusters were qualitatively different during the first and second time windows of reliable classifiability, substantiating the occurrence of distinct encoding stages. Change was particularly appreciable in the individual topographies (*SI Appendix*, Fig. S3 *A* and *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 (*SI Appendix*, Fig. S3) and found significant differences between brain activity patterns precisely distinctive for either labials, alveolars, or velars (*SI Appendix*, Fig. S4, in which 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 (29, 6).

An Invariant Code for Subsyllabic 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 (30, 31), we expected full generalization across voice genders. As explained in the introduction, successful generalization across vowels would be indicative of subsyllabic 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 (Fig. 3*A*, Table 1, and *SI Appendix*, Table S1). 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 (32). Remarkably, clear generalization across voices and vowels was also obtained for place (Fig. 3*B*). The time course of classification, with two distinct decodable periods, and its accuracy were comparable to those achieved in the initial analysis (Fig. 3*B*, Table 1, and *SI Appendix*, Table S1). Since the acoustic cues for place vary substantially with the context (33, 34), 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 (Fig. 3*C* and Table 1). Thus, the cross-decoding patterns observed so far demonstrate that syllables are not perceived holistically but broken down into subcomponents independently of the coarticulated vowel for consonants and consonantal features for vowels.

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/ versus /m/, /n/, /ŋ/) 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 at one featural context (e.g., manner classifiers were trained only on labials) and tested them on left-out data either within the same

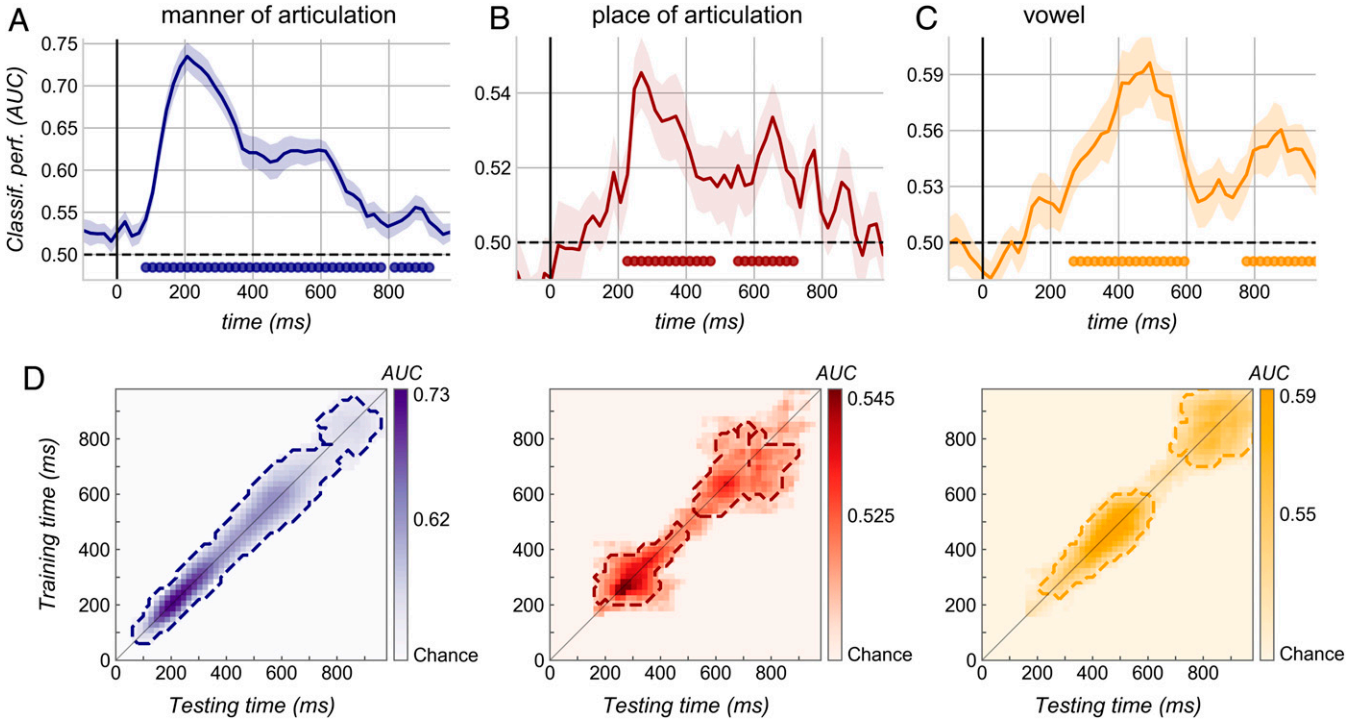

Fig. 2. Classification performances of estimators trained on single time windows (20 ms) along the ERP. (*Top*) Estimators are tested at the trained time sample. The shaded areas correspond to the SE (SEM) across subjects, the dotted black lines mark theoretical chance level, and the filled circles indicate significant scores (cluster-corrected Student's *t* test). (*A*) Performance of classifiers trained on manner distinctions: obstruents (/b/, /d/, and /g/) versus sonorants (/m/, /n/, and /ŋ/). (*B*) Performance of classifiers trained on place distinctions: labials (/b/ and /m/) versus alveolars (/d/ and /n/) versus velars (/g/ and /ŋ/). (*C*) Classification of vowel identities: /i/ versus /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. The dashed contours indicate significant clusters (manner: $p_{clust} = 0.0001$; place: $p_{clust} = 0.0001$ and 0.0028, vowel: $p_{clust} = 0.0097$ and $p_{clust} = 0.0108$).

condition (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 (Fig. 4*A*): 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 ∼450 ms, 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 (Fig. 4*A*). This finding suggests that a second phase of processing involved the grouping of multiple elementary dimensions into an integrated neural code. In other words, during this later time window, features were merged and no longer encoded as orthogonal, separately decodable dimensions.

Consonants 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 Figs. 2 and 3, 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 (Fig. 3*C*), 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 Fig. 4*B*, such a drop never occurred, suggesting that consonants and vowels were kept separated, at least until 1 s after syllable onset.

All the decoding results described above were further validated by the sanity check analyses illustrated in *SI Appendix*, Fig. S5 in which 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; and c) the nonarbitrariness of phonemes and phonetic features as relevant linguistic dimensions.

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" versus "bo" versus "di" versus "do" versus "gi" versus "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 (*SI Appendix*, Fig. S6 *A*, *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 (Fig. 4 *C*, *Left*, and *SI Appendix*, Fig. S6 *A*, *Bottom*) was explained by

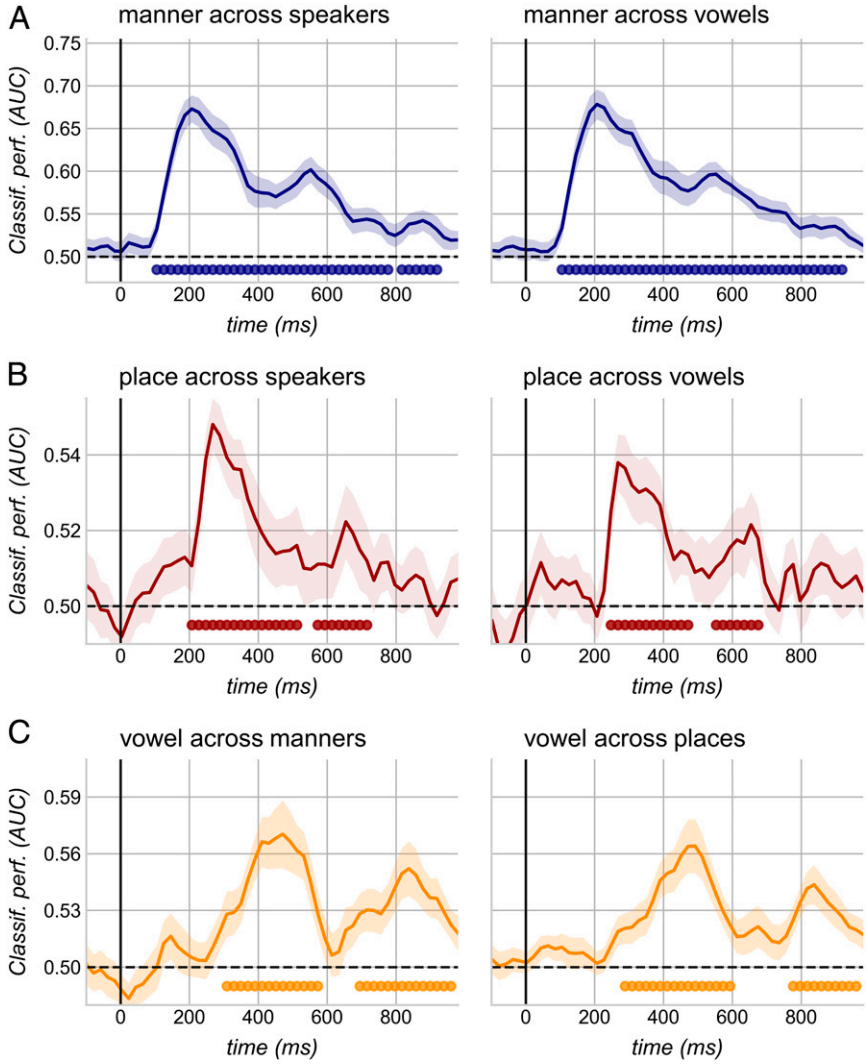

Fig. 3. Cross-condition decoding. (*A*, *Left*) Generalization of manner estimators across voice conditions: classifiers trained on syllables produced by one speaker are tested on stimuli uttered by the other speaker. (*Right*) Generalization of manner estimators across vowel conditions: classifiers 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). The shaded areas correspond to the SE (SEM) across subjects; the dotted black lines mark theoretical chance level. The filled circles indicate scores significantly above chance (exact *P* values are reported in Table 1). Performances from all possible training/test directions are averaged.

the isolation of either featural, consonant identity, and/or whole-syllable codes (Fig. 4 *C*, *Middle*) once vowel distinctions were entered as a variable of noninterest (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 to 380 ms: $p_{clust} = 0.017$). Crucially, consonant identity predicted the patterns of neural separability *only* later, between 500 and 700 ms (Fig. 4 *C*, *Right*; $p_{clust} = 0.006$). Lastly, the syllable regressor never reached significance (Fig. 4*C*). Complementing the decoding outcomes in Fig. 4 *A* and *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.

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 Fig. 2*D* and the corresponding informative activity patterns in *SI Appendix*, Figs. S3 and S4. 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 (35–38). 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

Table 1. Cross-condition decoding

Classes based on	Generalization across:	Time window (ms)	*p*-clust	Peak performance		
				Latency (ms)	Score	SD
Manner	Speakers	100 to 920	0.0001	200	0.673	0.079
	Vowels	100 to 920	0.0001	200	0.678	0.086
Place	Speakers	200 to 520	0.0001	260	0.548	0.035
		560 to 720	0.0014	640	0.522	0.047
	Vowels	240 to 480	0.0001	260	0.538	0.034
		540 to 680	0.006	640	0.522	0.042
Vowel	Speakers	260 to 580	0.0001	460	0.561	0.078
		760 to 920	0.0002	800	0.554	0.052
	Manners	300 to 580	0.0001	460	0.57	0.08
		680 to 960	0.0001	820	0.552	0.067
	Places	280 to 600	0.0001	480	0.564	0.082
		760 to 960	0.0001	820	0.544	0.066

Statistical description of the decoding performances shown in Fig. 3.

sounds in phonetic terms (35, 39, 38). In this scenario, canonical babbling, which signals the beginning of a fairly controlled articulation around 6 to 8 mo of age (40), represents an important milestone, while infants in the pre-babbling phase are thought to rely on refined but domain-general auditory mechanisms (41). It follows that, according to these widely accepted views, the primitive units for speech processing consist of spectro-temporally 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., ref. 42), our finding of neural codes for consonant identity complements adult data (43) in corroborating the reality of the phoneme as a relevant entity for the cortical encoding of speech (44).

Second, our generalization approach, involving the comparison of decoding performances within and across phonetic domains, disclosed the existence of a preliminary phase in which 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 a first encoding stage. To start with, we carefully selected the stimuli to avoid any trivial difference, for instance, in consonant duration (*SI Appendix, Stimuli Construction*). Importantly, we opted for the dimensions of place and manner because the consistency of their acoustic correlates across contexts is largely different. Furthermore, the experimental stimuli were appositely chosen to push the variability of place cues at the maximum [e.g., /i/ versus /o/, situated at opposite corners of the vowel diagram, accentuated the spectro-acoustical inconsistency of place cues due to coarticulatory phenomena (33)]. Such a prioris were confirmed by our inspection of the auditory spectrograms (*SI Appendix*, Fig. S1), in which 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 spectro-temporal 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 (4, 6, 29, 43, 45) and compatible with the decomposition into distinctive features postulated by linguists (3). 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 to 400 ms relative to phoneme onset (45). However, the experimental design did not allow to explore the functional significance of such evolving activity patterns. Cortical recordings in adults have uncovered how distinct electrodes encode different dimensions of the speech signal (4), but they could primarily capture 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, when applied to fMRI data, a multivariate decoding procedure equatable with that proposed here disclosed feature-specific responses in various areas of the adult temporal lobe (6). 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 (6) constitutes the first step of a perceptual process (45) leading to phoneme identity computation. Such a process creates a structured and highly generalizable space that is robust to surface variability across speakers and coarticulatory contexts.

A factorized representational mechanism was previously discovered in the monkey face patch system (2). As outlined for the visual domain, such a decomposition strategy applied to speech is more parsimonious, efficient, and flexible than exemplar coding (e.g., refs. 46 and 47). 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 (48). In this context, the vectorized system we propose discards the irrelevant variability to organize the input according to phonetic criteria; such perceptual reorganization turns up those subtle phonetic differences that define word's meaning. Importantly, in order to discover words, infants must cope not only with

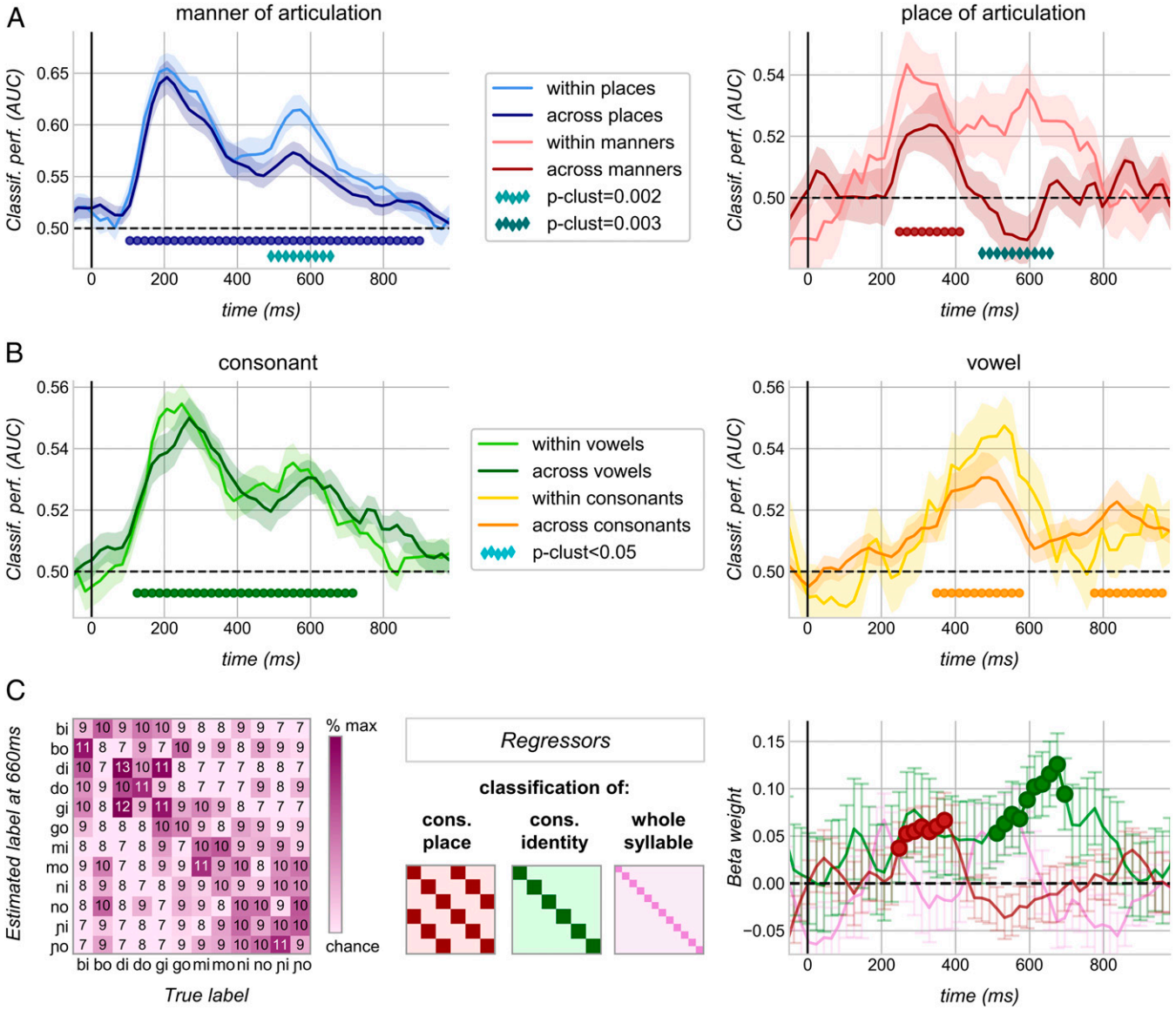

Fig. 4. Orthogonal feature codes are merged into phoneme identities at a late stage of processing. (*A*) Time-resolved performance of estimators trained on a single phonetic feature (e.g., manner estimators trained on labials: /b/ versus /m/). In light colors: classification within the trained condition (e.g., test on labials); in darker colors: performance at novel phonetic contexts (e.g., test on alveolars: /d/ versus /n/ and velars: /g/ versus /ŋ/). 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 significant generalization across contexts (100 to 900 ms: $p_{clust} = 0.0001$ for manner; 240 to 420 ms: $p_{clust} = 0.001$ for place). The diamonds indicate higher performance within as compared to across conditions (exact time window of significance for manner: 480 to 640 ms; for place: 460 to 660 ms). (*B*, *Left*) Performance of estimators trained on discriminating all consonants (/b/ versus /d/ versus /g/ versus /m/ versus /n/ versus /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 to 900 ms, $p_{clust} = 0.0001$; vowel classifiers: 340 to 560 ms, $p_{clust} = 0.0001$ and 760 ms onwards, $p_{clust} = 0.0002$). (*C*, *Left*) Example of a neural confusion matrix at time *t* (660 ms) obtained with a 12-class (syllables) decoding problem (average across subjects). The 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). The darker colors correspond to the values 33.3, 50, and 100%, respectively, and the 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 Student's *t* test). The vertical lines correspond to the SEM. To enhance clarity, the remaining predictors (i.e., manner and vowel discrimination) and the relative beta weights are illustrated in *SI Appendix*, Fig. S6*B*.

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 (49). 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 mo: they prefer to listen to sequences that are phonotactically legal in their native language (50, 51) and use their phonotactic knowledge to

find word boundaries in continuous speech (52). 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 (53). Lastly, a featural encoding of speech is consistent with the documented ability of young infants to use phonetic details in word-referent mapping (54, 55).

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 (56). 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 (57), infants are known to exploit the "division of labor" between consonants and vowels already by the age of 12 mo (58). The inclusion of different syllabic structures in future experimental paradigms will bring further insights on this matter (for example, 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 thus 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 wk, the age of our subjects, vocal production is very limited (41). Strikingly, even preterm neonates can detect a place of articulation change ("ba" versus "ga") at 6 mo 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 (59, 60). 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 (61). These primitive behaviors could provide a primordial knowledge of the shape and configurability of the upper vocal tract (62), and, combined with sound exposure, they might foster an integrative/multimodal representational space for speech before the onset of canonical babbling. Coherently with this conjecture, a recent study in 3 mo 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 (63). Although multimodal speech processing appears from an early age (31), 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 (25, 64). When conceiving the brain as an information processing system based on population coding (65), pattern-information analyses are likely to have considerable functional significance, especially in comparison to more classical activation-based approaches (25, 64). Furthermore, our choice of linear (as opposed to nonlinear) classifiers ensures the biological plausibility of our conclusions (66). 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 mo olds could successfully learn a phonotactic rule shaping VC pairings on the basis of featural classes (i.e., "nasal vowels are always followed by fricatives and oral vowels by stop consonants") (67). Moreover, a recent ERP study found that when exposed to syllables varying in their vocalic constituents, 3-mo-old infants could learn to pair consonants with visual shapes and generalize this pairing to a new vocalic context, demonstrating that subsyllabic representations are already operational at this age (20). 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 (Fig. 4*C* and *SI Appendix*, Fig. S6). Namely, in the 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 (68). 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.

Materials and Methods

Participants. A total of 25 full-term, normal-hearing infants (12 females, 13 males) coming from a French-speaking environment were tested between 12 and 14 wk after birth (mean age = 12 wk and 6 d). 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 in *EEG Recording and Data Preprocessing*), technical problems during data collection ($n = 3$), or aberrant global field power (GFP) in the average of all syllable-related potentials (i.e., peak GFP < 4 μV, $n = 4$). The protocol was approved by the regional ethical committee for biomedical research (Comité de Protection des Personnes Region Centre Ouest 1). Parents gave their written informed consent before starting the experiment.

Stimuli. Stimuli consisted of 120 speech sounds constructed upon six consonants: /b/, /d/, /g/, /m/, /n/, and /ɲ/. These consonants were selected to cover two manner features, that is, obstruent (/b/, /d/, and /g/) and sonorant (/m/, /n/, and /ɲ/), and three places of articulation, that is, labial (/b/ and /m/), alveolar (/d/ and /n/), and palatal–velar (/g/ and /ɲ/, 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 × 3 place × 2 vowel × 2 voice factor design (i.e., 24 subconditions). To increase acoustic variability (and extend the external validity of our measurements), speakers were asked to repeat the same tokens several times while changing their intonation. For every subcondition, we selected five 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 spectro-temporal profiles (3 consonants × 2 vowels × 2 voices × 5 utterances); similarly, each place of articulation was presented in 40 (2 consonants × 2 vowels × 2 voices × 5 utterances) spectro-temporal versions.

Further details are provided in *SI Appendix, Stimuli Construction*.

Procedure. Subjects were tested in a soundproof Faraday cage equipped with a computer screen and loudspeakers on the top. Infants were held by a caregiver, and 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 interstimulus interval between 600 and 1,000 ms. 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 3,136 tokens (corresponding to ∼63 min of listening time) or as soon as infants became restless.

EEG Recording and Data Preprocessing. The 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 (Electrical Geodesics, Inc.) referenced to the vertex. The sensor layout of this prototype diverges from the classical geodesic 128 locations partitioning (69) in that 20 of the standard temporal positions are covered by two tight grids of sensors (70 electrodes on each side, organized in hexagonal pods) with no sponge inserts (*SI Appendix*, Fig. S2). Electrodes are made of carbon fibers embedded within a plastic (ABS) substrate and coated with silver chloride.

Artifact Detection and Correction. Data preprocessing was conducted through custom-made MATLAB scripts based on the EEGLAB toolbox 14.0 (70). While following the main preprocessing steps normally used in developmental studies, we introduced some modifications inspired by efforts carried to improve adult data quality (71, 72). 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 interindividual 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 (*SI Appendix*, Fig. S2) and high sampling rate.

As a first step, EEG recordings were band-pass filtered (0.5 to 40 Hz), 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 following: the voltage amplitude and its first derivative, the variance across a 500-ms long moving time window, and the fast-running average and the deviation between the fast- and the slow-running averages within a 500-ms 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 $\pm$ $n \times$ IQ, where IQ is the interquartile range of the distribution). Two additional algorithms identified whether the power within the 0- to 10-Hz band was excessively low or within 20 to 40 Hz 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 50 ms with a percentage of rejected channels superior to 30% or beyond 2IQ from the third 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 third 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 corrections were applied. When the rejected segments had a very short duration (50 ms 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, 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 (73). 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 (0.5 Hz) 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 200 ms before and ending 1,400 ms 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, that is, when their mean or maximum distance from the average response was an outlier in the distribution (>3rd quartile + 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 number of trials, we included subjects with a minimum of 40 epochs/subcondition. 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 subcondition (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 20 Hz, mathematically re-referenced to the mean of all channels, and down sampled (with a moving average of two time points) to 250 Hz. All the main analyses (decoding) were carried at the single-trial level. Nonetheless, epochs were also averaged per either subcondition or manner/place condition in order to examine evoked responses (ERPs, e.g., *SI Appendix*, Fig. S4*C*).

Decoding. Multivariate pattern analyses were conducted within subject, relying on the Scikit-Learn (74) and MNE (75, 76) Python packages. To decode *in time*, epochs were divided into 60 consecutive windows of 20 ms (from −200 to 1,000 ms relative to stimulus onset), each corresponding to a matrix with the shape n channels × 5 samples (sampling rate = 250 Hz, 5 samples = 20 ms). 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 labeled as belonging to either the category of "obstruent" or to the category of "sonorant" depending on whether /b/, /d/, /g/ or /m/, /n/, /ŋ/ exemplars were presented. To decode the place of articulation, y comprised three classes: "labial" (/b/ and /m/), "alveolar" (/d/ and /n/), and "velar" (/g/ and /ŋ/). 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 and b) that 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 versus 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 the training and the test set for a given run were defined, we applied a "micro-averaging" procedure, a strategy previously used on adults with the same purpose (77). 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 one to three randomly picked trials from the most numerous class(es).

Next, following the z-scoring of each feature (i.e., channel and time point across trials), a L1-norm regularized logistic regression (78) 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 (79)] 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-versus-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. The outcome of each algorithm was a vector 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-versus-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, the outcomes from all cross-validation runs were averaged.

As a proof of concept, the main decoding analyses were performed with two additional algorithms: L1-norm regularized linear support vector machine (78) and linear discriminant analysis. 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 (80). Importantly, we restricted our alternatives to linear classifiers to make sure that the algorithms focused on explicit neural codes (66). Besides 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', that is, every 20 ms from 200 ms prior to 1,000 ms after syllable onset. Such a procedure was performed within the cross-validation so that the 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 (27). 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 (27).

Generalization across Conditions. We examined the consistency of information used by classifiers in different harmonic and coarticulatory 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/, and /g/ versus /m/, /n/, and /ɲ/), place contrasts (/b/ and /m/ versus /d/ and /n/ versus /g/ and /ɲ/), and vowel contrasts (/i/ versus /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 coarticulations 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" versus "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 the 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 versus alveolars versus 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/ versus /d/ versus /g/ versus /m/ versus /n/ versus /ɲ/ when they were coarticulated with /i/) and tested them within the same and across the alternative coarticulatory context (e.g., classify consonant identity among "bo," "do," "go," "mo," "no," and "ɲo"; note that for this schema, as for place classification, we adopted a "one-versus-rest" approach). 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 two micro-averaged trials/class at test and at the same time maximize the number 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 eight epochs. Apart from these modifications, the decoding procedures resembled those described above.

Neural Syllable Confusion and Multiple Regression Analysis. For this section, we first built a 12-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" versus "bo" versus "di" versus "do" versus "gi," etc.). We adopted a "one-versus-rest" approach and used the same preprocessing 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 in which 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 (Fig. 4 *C*, *Left*, and *SI Appendix*, Fig. S6 *A*, *Bottom*). The diagonal of these confusion matrices depicts class-wise accuracy, with theoretical chance being at 8.3% (*SI Appendix*, Fig. S6 *A*, *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 (81) 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 (Fig. 4 *C*, *Middle*, and *SI Appendix*, Fig. S6 *B*, *Top*). Concerning the matrix modeling 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.

Statistical Analysis. To calculate statistics, we performed second-level tests across subjects employing the MNE dedicated functions. Following the example in ref. 82, 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, and c) whether multiple regression beta weights were higher than zero using one-sample cluster-based permutation t tests (83), 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 bidimensional. 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 10,000 random sign flips of the observed outcomes. A cluster was considered as significant when its P value was below 0.05.

Data Availability. Anonymized EEG infant data and supporting material are available in Zenodo: https://doi.org/10.5281/zenodo.4579401.

ACKNOWLEDGMENTS. This research was supported by grants from the Fondation NrJ, Fondation Bettencourt, and the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement 695710). We thank Don Tucker and Amy Rowland (Electrical Geodesics, Inc. and University of Oregon) for designing the 256 electrodes net and Bahar Khalighinejad for providing help with auditory spectrogram estimation. We also thank Stanislas Dehaene, Yair Lakretz, and Christophe Pallier for constructive feedback and suggestions. Finally, we are grateful for the essential mentorship provided by Sébastien Marti. His smile, kindness, and friendship remain in our hearts.

1. T. E. J. Behrens *et al*., What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
2. L. Chang, D. Y. Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013–1028.e14 (2017).
3. M. Halle,*From Memory to Speech and Back: Papers on Phonetics and Phonology 1954-2002* (Walter de Gruyter, 2013).
4. N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).

5. J. S. Arsenault, B. R. Buchsbaum, Distributed neural representations of phonological features during speech perception. *J. Neurosci.* **35**, 634–642 (2015).
6. J. M. Correia, B. M. B. Jansma, M. Bonte, Decoding articulatory features from fMRI responses in dorsal speech regions. *J. Neurosci.* **35**, 15015–15025 (2015).
7. P. D. Eimas, E. R. Siqueland, P. Jusczyk, J. Vigorito, Speech perception in infants. *Science* **171**, 303–306 (1971).
8. J. Bertoncini, R. Bijeljac-Babic, S. E. Blumstein, J. Mehler, Discrimination in neonates of very short CVs. *J. Acoust. Soc. Am.* **82**, 31–37 (1987).
9. P. D. Eimas, J. L. Miller, Discrimination of information for manner of articulation. *Infant Behav. Dev.* **3**, 367–375 (1980).
10. G. Dehaene-Lambertz, M. Pena, Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport* **12**, 3155–3158 (2001).
11. P. W. Jusczyk, D. B. Pisoni, J. Mullennix, Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition* **43**, 253–291 (1992).
12. P. D. Eimas, J. L. Miller, Contextual effects in infant speech perception. *Science* **209**, 1140–1141 (1980).
13. J. L. Miller, P. D. Eimas, Studies on the categorization of speech by infants. *Cognition* **13**, 135–165 (1983).
14. A. Fló *et al*., Newborns are sensitive to multiple cues for word segmentation in continuous speech. *Dev. Sci.* **22**, e12802 (2019).
15. P. W. Jusczyk, C. Derrah, Representation of speech sounds by young infants. *Dev. Psychol.* **23**, 648–654 (1987).
16. J. Bertoncini, R. Bijeljac-Babic, P. W. Jusczyk, L. J. Kennedy, J. Mehler, An investigation of young infants' perceptual representations of speech sounds. *J. Exp. Psychol. Gen.* **117**, 21–33 (1988).
17. R. Bijeljac-Babic, J. Bertoncini, J. Mehler, How do 4-day-old infants categorize multisyllabic utterances? *Dev. Psychol.* **29**, 711–721 (1993).
18. O. Räsänen, G. Doyle, M. C. Frank, Pre-linguistic segmentation of speech into syllable-like units. *Cognition* **171**, 130–150 (2018).
19. K. Mersad, G. Dehaene-Lambertz, Electrophysiological evidence of phonetic normalization across coarticulation in infants. *Dev. Sci.* **19**, 710–722 (2016).
20. K. Mersad, C. Kabdebon, G. Dehaene-Lambertz, Explicit access to phonetic representations in 3-month-old infants. *Cognition* 10.1016/j.cognition.2021.104613 (2021).
21. K. N. Stevens, *Acoustic Phonetics* (MIT Press, 2000).
22. R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, Speech recognition with primarily temporal cues. *Science* **270**, 303–304 (1995).
23. R. Smits, L. ten Bosch, R. Collier, Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluation. *J. Acoust. Soc. Am.* **100**, 3865–3881 (1996).
24. C. A. Fowler, Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Percept. Psychophys.* **55**, 597–610 (1994).
25. M. N. Hebart, C. I. Baker, Deconstructing multivariate decoding for the study of brain function. *Neuroimage* **180** (Pt A), 4–18 (2018).
26. N. Kriegeskorte, P. K. Douglas, Interpreting encoding and decoding models. *Curr. Opin. Neurobiol.* **55**, 167–179 (2019).
27. J.-R. King, S. Dehaene, Characterizing the dynamics of mental representations: The temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).
28. M. G. Stokes, M. J. Wolff, E. Spaak, Decoding rich spatial information with high temporal resolution. *Trends Cogn. Sci.* **19**, 636–638 (2015).
29. E. F. Chang *et al*., Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* **13**, 1428–1432 (2010).
30. E. Formisano, F. De Martino, M. Bonte, R. Goebel, "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* **322**, 970–973 (2008).
31. D. Bristow *et al*., Hearing faces: How the infant brain matches the face it sees with the speech it hears. *J. Cogn. Neurosci.* **21**, 905–921 (2009).
32. J. Hillenbrand, Perceptual organization of speech sounds by infants. *J. Speech Lang. Hear. Res.* **26**, 268–282 (1983).
33. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, Perception of the speech code. *Psychol. Rev.* **74**, 431–461 (1967).
34. M. F. Dorman, M. Studdert-Kennedy, L. J. Raphael, Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Percept. Psychophys.* **22**, 109–122 (1977).
35. P. K. Kuhl *et al*., Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 979–1000 (2008).
36. G. Westermann, E. Reck Miranda, A new model of sensorimotor coupling in the development of speech. *Brain Lang.* **89**, 393–400 (2004).
37. J.-L. Schwartz, A. Basirat, L. Ménard, M. Sato, The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception. *J. Neurolinguist.* **25**, 336–354 (2012).
38. A. Vilain, M. Dole, H. Lœvenbruck, O. Pascalis, J.-L. Schwartz, The role of production abilities in the perception of consonant category in infants. *Dev. Sci.* **22**, e12830 (2019).
39. P. K. Kuhl, R. R. Ramírez, A. Bosseler, J.-F. L. Lin, T. Imada, Infants' brain responses to speech suggest analysis by synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 11238–11245 (2014).
40. J. M. van der Stelt, F. J. Koopmans-van Beinum, "The onset of babbling related to gross motor development" in *Precursors of Early Speech: Proceedings of an International Symposium Held at The Wenner-Gren Center*, Stockholm, September 19–22, 1984, Wenner-Gren Center International Symposium Series, B. Lindblom, R. Zetterström, Eds. (Palgrave Macmillan UK, 1986), pp. 163–173.
41. P. K. Kuhl, Early language acquisition: Cracking the speech code. *Nat. Rev. Neurosci.* **5**, 831–843 (2004).
42. Y. Oganian, E. F. Chang, A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci. Adv.* **5**, eaay6279 (2019).
43. Q. Zhang *et al*., Deciphering phonemes from syllables in blood oxygenation level-dependent signals in human superior temporal gyrus. *Eur. J. Neurosci.* **43**, 773–781 (2016).
44. N. Kazanina, J. S. Bowers, W. Idsardi, Phonemes: Lexical access and beyond. *Psychon. Bull. Rev.* **25**, 560–585 (2018).
45. B. Khalighinejad, G. Cruzatto da Silva, N. Mesgarani, Dynamic encoding of acoustic features in neural responses to continuous speech. *J. Neurosci.* **37**, 2176–2185 (2017).
46. R. Port, How are words stored in memory? Beyond phones and phonemes. *New Ideas Psychol.* **25**, 143–170 (2007).
47. R. F. Port, Rich memory and distributed phonology. *Lang. Sci.* **32**, 43–55 (2010).
48. G. C. Rost, B. McMurray, Speaker variability augments phonological processing in early word learning. *Dev. Sci.* **12**, 339–349 (2009).
49. I. Darcy, F. Ramus, A. Christophe, K. Kinzler, E. Dupoux, "Phonological knowledge in compensation for native and non-native assimilation" in *Variation and Gradience in Phonetics and Phonology*, F. Kügler, C. Féry, R. van de Vijver, Eds. (Mouton de Gruyter, 2009), pp. 265–310.
50. A. D. Friederici, J. M. I. Wessels, Phonotactic knowledge of word boundaries and its use in infant speech perception. *Percept. Psychophys.* **54**, 287–295 (1993).
51. P. W. Jusczyk, A. D. Friederici, J. M. I. Wessels, V. Y. Svenkerud, A. M. Jusczyk, Infants' sensitivity to the sound patterns of native language words. *J. Mem. Lang.* **32**, 402–420 (1993).
52. S. L. Mattys, P. W. Jusczyk, Phonotactic cues for segmentation of fluent speech by infants. *Cognition* **78**, 91–121 (2001).
53. J. R. Saffran, E. D. Thiessen, Pattern induction by infant language learners. *Dev. Psychol.* **39**, 484–494 (2003).
54. D. Swingley, R. N. Aslin, Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychol. Sci.* **13**, 480–484 (2002).
55. C. T. Fennell, S. R. Waxman, What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Dev.* **81**, 1376–1383 (2010).
56. M. Nespor, M. Peña, J. Mehler, *On the Different Roles of Vowels and Consonants in Speech Processing and Language Acquisition* (Lingue E Linguaggio, 2003).
57. J. M. Toro, M. Nespor, J. Mehler, L. L. Bonatti, Finding words and rules in a speech stream: Functional differences between vowels and consonants. *Psychol. Sci.* **19**, 137–144 (2008).
58. J.-R. Hochmann, S. Benavides-Varela, M. Nespor, J. Mehler, Consonants and vowels: Different roles in early language acquisition. *Dev. Sci.* **14**, 1445–1458 (2011).
59. M. Mahmoudzadeh *et al*., Syllabic discrimination in premature human infants prior to complete formation of cortical layers. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 4846–4851 (2013).
60. M. Mahmoudzadeh, F. Wallois, G. Kongolo, S. Goudjil, G. Dehaene-Lambertz, Functional maps at the onset of auditory inputs in very early preterm human neonates. *Cereb. Cortex* **27**, 2500–2512 (2016).
61. D. K. Oller *et al*., Preterm and full term infant vocalization and the origin of language. *Sci. Rep.* **9**, 14734 (2019).
62. D. Choi, P. Kandhadai, D. K. Danielson, A. G. Bruderer, J. F. Werker, Does early motor development contribute to speech perception? *Behav. Brain Sci.* **40**, e388 (2017).
63. D. Choi, G. Dehaene-Lambertz, M. Peña, J. F. Werker, Neural indicators of articulator-specific sensorimotor influences on infant speech perception. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2025043118 (2021).
64. N. Kriegeskorte, P. Bandettini, Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage* **38**, 649–662 (2007).
65. S. Panzeri, J. H. Macke, J. Gross, C. Kayser, Neural population coding: Combining insights from microscopic and mass signals. *Trends Cogn. Sci.* **19**, 162–172 (2015).
66. N. Kriegeskorte, Pattern-information analysis: From stimulus decoding to computational-model testing. *Neuroimage* **56**, 411–421 (2011).
67. A. Seidl, A. Cristià, A. Bernard, K. H. Onishi, Allophonic and phonemic contrasts in infants' learning of sound patterns. *Lang. Learn. Dev.* **5**, 191–202 (2009).
68. P. W. Jusczyk, "Early research on speech perception" in *The Discovery of Spoken Language* (MIT Press, 2000), pp. 43–71.
69. D. M. Tucker, Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalogr. Clin. Neurophysiol.* **87**, 154–163 (1993).
70. A. Delorme, S. Makeig, EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
71. M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, A. Gramfort, Autoreject: Automated artifact rejection for MEG and EEG data. *Neuroimage* **159**, 417–429 (2017).
72. A. Mognon, J. Jovicich, L. Bruzzone, M. Buiatti, ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* **48**, 229–240 (2011).
73. F. Perrin, J. Pernier, O. Bertrand, J. F. Echallier, Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neurophysiol.* **72**, 184–187 (1989).
74. F. Pedregosa *et al*., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
75. A. Gramfort *et al*., MNE software for processing MEG and EEG data. *Neuroimage* **86**, 446–460 (2014).
76. A. Gramfort *et al*., MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7**, 267 (2013).
77. T. Grootswagers, S. G. Wardle, T. A. Carlson, Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J. Cogn. Neurosci.* **29**, 677–697 (2017).
78. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
79. A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance" in *Proceedings of the twenty-first international conference on Machine learning* (Banff, Canada, July, 2004).
80. O. Ledoit, M. Wolf, Honey, I shrunk the sample covariance matrix. *J. Portf. Manag.* **30**, 110–119 (2003).
81. N. Kriegeskorte, R. A. Kievit, Representational geometry: Integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).
82. J.-R. King, N. Pescetelli, S. Dehaene, Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron* **92**, 1122–1134 (2016).
83. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).