# Speech perception at birth: The brain encodes fast and slow temporal information

Laurianne Cabrera, Judit Gervain

# Speech perception at birth:

# the brain encodes fast and slow temporal information

**Laurianne Cabrera[1,2*] and Judit Gervain[1,2]**

[1] Integrative Neuroscience and Cognition Center
Université de Paris, UFR Biomédicale
45 rue des Saints Pères
75006, Paris, France

[2] Integrative Neuroscience and Cognition Center
CNRS
45 rue des Saints Pères
75006, Paris, France

* Corresponding author. Email: laurianne.cabrera@parisdescartes.fr

*Paper published in Science Advances, 22 Jul 2020, doi: 10.1126/sciadv.aba7830*

**Abstract**

Speech perception is constrained by auditory processing. Although at birth, infants have an immature auditory system and limited language experience, they show remarkable speech perception skills. To assess neonates' ability to process the complex acoustic cues of speech, we combined near-infrared spectroscopy (NIRS) and electroencephalography (EEG) to measure brain responses to syllables differing in consonants. The syllables were presented in three conditions preserving (i) original temporal modulations of speech (both amplitude and frequency modulations, AM/FM), (ii) both fast and slow AM, but not FM, or (iii) only the slowest AM (< 8 Hz). EEG responses indicate that neonates are able to encode consonants in all conditions, even without the fast temporal modulations, similarly to adults. Yet, the fast and slow AM activate different neural areas, as shown by NIRS. Thus, the *immature* human brain is already able to decompose the acoustic components of speech, laying the foundations of language learning.

**Keywords:** temporal envelope, amplitude modulations, EEG, fNIRS, newborns, speech perception, phoneme discrimination, auditory development

## Introduction

Speech perception requires efficient auditory mechanisms to track subtle differences in the complex combination of spectral and temporal information differentiating linguistic contrasts. Although infants have an immature peripheral and central auditory system (1, 2), they show exquisite speech perception abilities from birth (3–6). How they can achieve this, and whether they rely on the same acoustic information as adults remains unknown. The present study investigates whether, and if yes, how newborn infants use the temporal information in the speech signal to discriminate phonemes.

Temporal information plays an essential role in speech perception in adults. Speech is mainly conveyed to the brain by the basilar membrane in the cochlea, the inner ear, which encodes the temporal modulations of the speech signal in different frequency regions, or bands (7). Within each frequency band, the temporal properties are extracted at two time scales: the amplitude modulation cues (AM), also called temporal envelope, corresponding to the relatively slow variations in amplitude over time, and the frequency modulation (FM) cues, also called temporal fine structure, corresponding to the variations in instantaneous frequency close to the center frequency of the frequency band.

This temporal decomposition is observed at the cortical level in adults. Previous studies measuring brain activation for non-speech sounds modulated in amplitude at different rates showed predominant cortical responses to the lowest AM frequencies (4-8 Hz) and hemispheric specialization in temporal envelope coding, as well as a difference in the time course of activations between low (< 16 Hz) and fast (< 128 Hz) AM rates (8, 9). For speech sounds, the debate about the hemispheric specialization to different acoustic properties of the speech signal is ongoing (10), but it is usually assumed that fast temporal modulation is preferentially processed by the left auditory cortex, while slow temporal modulation and/or spectral modulation is processed by the right temporal cortex (11, 12).

These temporal modulations also play different roles for speech perception, as different rates of modulations convey different linguistic information. A wealth of studies in psychoacoustics showed that the slowest envelope cues (under 16 Hz) play a primary role in the identification of consonants, vowels and words in speech presented in quiet (13, 14). Faster envelope cues (closer to the fundamental frequency rate of the voice) and the temporal fine structure play a more important role in perceiving pitch, which in turn contributes to the comprehension of speech in noise as well as of linguistic units heavily dependent on pitch information, such as lexical tone (7, 15, 16).

These neuroimaging and behavioral studies only focused on adult listeners who have a *mature* auditory and linguistic systems. But the auditory system takes years to mature. It is thus possible that the immature auditory system of infants decodes sound differently than that of adults. If so, this has important consequences for language development, which also unfolds during the first years of life, as its auditory input would thus differ from what adults perceive when processing speech. Currently, we have very little knowledge about how the youngest learners perceive the acoustic details of speech. The current study aims to fill this gap.

The few existing behavioral studies with infants (17–21) suggest that 6-month-olds might weigh modulation cues differently than adults. Indeed, even though 6-month-old French infants, like adults, are able to use the speech envelope to discriminate consonants based on voicing such as in /aba/-/apa/, and place of articulation such as in /aba/-/ada/ in quiet, they require more time to habituate to speech sounds containing only envelope cues below 16 Hz than to speech sounds preserving faster modulations. Moreover, 3-month-old infants and adults do not rely similarly on the fast and slow AM cues in quiet and in noise (21). Infants require the fast AM cues (> 8 Hz) in both quiet and noise when discriminating plosive consonants, while the slowest AM cues (< 8Hz) are sufficient for adults in quiet but

they also need the faster modulations in noise. These results suggest that fast envelope cues may be important for consonant perception in infants even in quiet.

The neural basis of the auditory processing of temporal modulations is still not well understood in infancy and has never been investigated using complex acoustic signals such as speech. One study exploring newborns' neural responses to non-speech sounds with different temporal structures suggested that, as adults, newborns show different neural responses to slowly (~ 3-8 Hz) and fast modulated signals (~ 40 Hz), with greater bilateral temporal activations for the later (22), as measured by near-infrared spectroscopy (NIRS), although the auditory evoked potentials recorded using electroencephalography (EEG) were not different. No study has directly compared newborns' perception of the slow versus fast temporal modulations of speech. The interaction between auditory mechanisms and speech perception at early stages of human development, therefore, remain to a large extent unexplored.

To determine how newborn infants, who have little experience with their native language and an immature auditory system, process the temporal acoustic cues of speech to perceive consonants, we used two hitherto rarely combined approaches. We combined a vocoder manipulation of speech with brain imaging in order to test how newborns process and perceive the temporal modulations in speech, essential for speech intelligibility in adulthood. Vocoders are powerful speech analysis and synthesis tools that can selectively manipulate the spectro-temporal properties of sound (14). We used a vocoder to selectively manipulate simple C(onsonant)-V(owel) syllables in three conditions: (i) the 'intact' condition preserved both the temporal envelope and the temporal fine structure, closely matching the original signal and serving as a control for the vocoding manipulations, (ii) the 'fast' condition preserved both the fast and the slow envelope components, thus retaining some voice-pitch and formant transition information (< 500 Hz), but suppressed the temporal fine structure, while (iii) the 'slow' condition only preserved the slowest temporal envelope

(< 8 Hz), retaining mainly the modulations related to syllables (23). A group of 2-day-old, healthy, full-term French neonates heard syllables differing in their consonants (/pa/-/ta/) in these three conditions (Fig. 1). We recorded newborns' brain responses to these speech sounds combining EEG and NIRS (Fig. 2) to assess the electrophysiological brain response and its metabolic correlates, respectively. Coupling these two techniques, which has rarely been done before in young infants (22, 24), has the unique advantage of providing both accurate spatial localization and high temporal resolution.

While newborns were lying quietly in their hospital cribs, the syllables were presented to them through loudspeakers in long stimulation blocks (30 sec) with 6 blocks per condition (Fig. 3), satisfying the temporal requirements of the slow hemodynamic response measured by NIRS. The intact sound condition was always played last in order to avoid priming, while the order of the slow and fast conditions was counterbalanced across babies. Each block contained 25 syllables, out of which 20 were standard syllables (e.g. /pa/) and 5 were deviants (e.g. /ta/), allowing an event-related assessment of the responses to individual syllables within blocks similarly to the classical oddball or mismatch design in EEG studies (25). Thus, each block of stimulation comprised a ratio of 80-20% of standard and deviant sounds. The only difference between standard and deviant sounds is the consonant at the onset of the syllable. The first 5 sounds were always standards to allow expectations about the standard to build up. The standard and deviant syllables were counterbalanced across babies. This design allowed us to address two research questions. First, whether overall the newborn brain processes the slow, fast and intact conditions similarly or differently, as indexed by the comparison of the hemodynamic responses to the three conditions. Second, whether newborns can successfully discriminate consonants on the basis of the temporal acoustic cues present in each of the conditions, as indexed by the event-related EEG response to the standard vs. deviant syllables

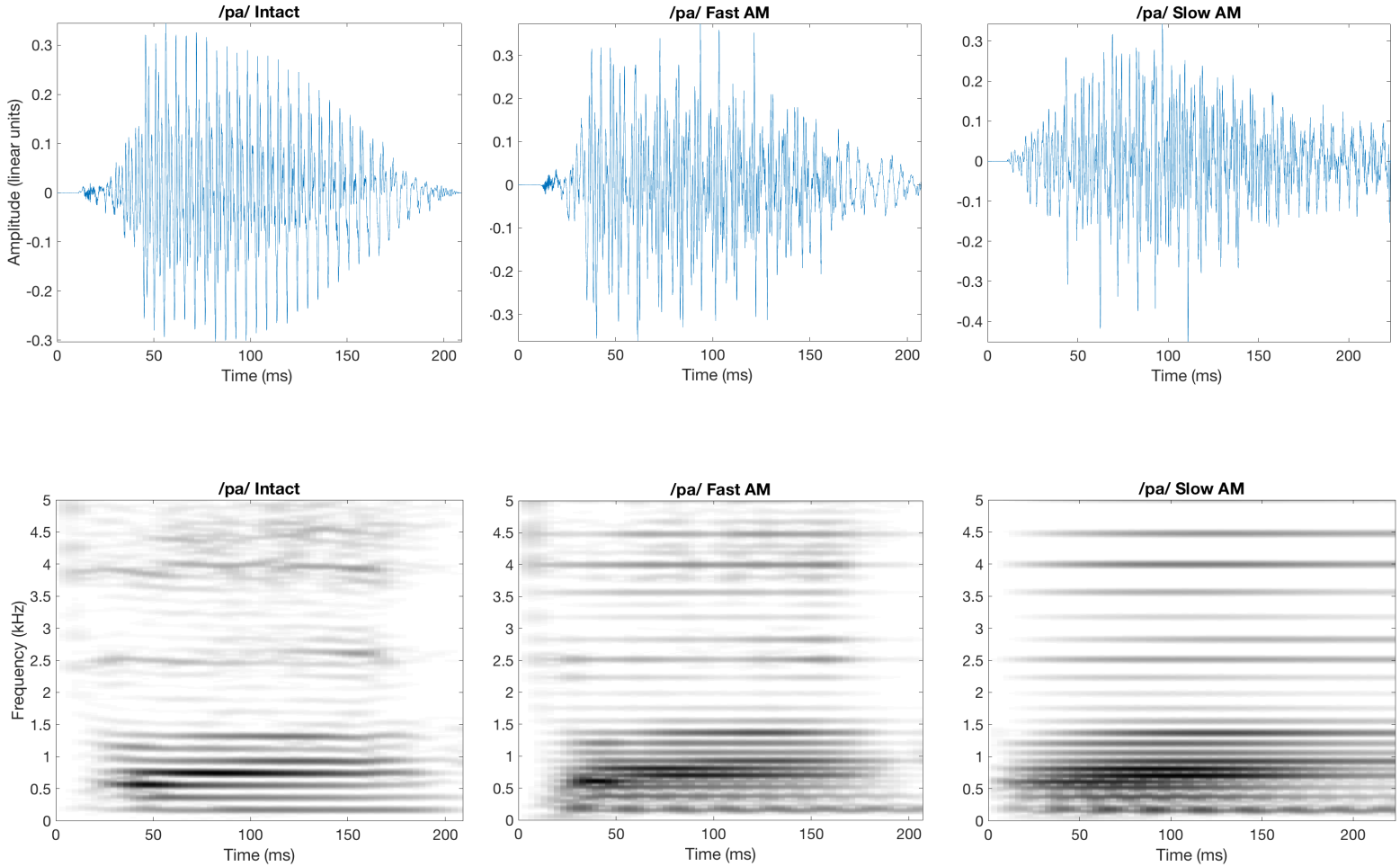


**Fig. 1. Waveforms and spectrograms of one syllable exemplar**. Waveforms (upper lines) and spectrograms (lower lines) of /pa/ filtered in the condition Intact (AM + FM) on the left, Fast AM (AM < $ERB_N/2$) on the middle and the condition Slow AM (AM < 8 Hz) on the right. AM indicates amplitude modulation and ERB equivalent rectangular bandwidth.

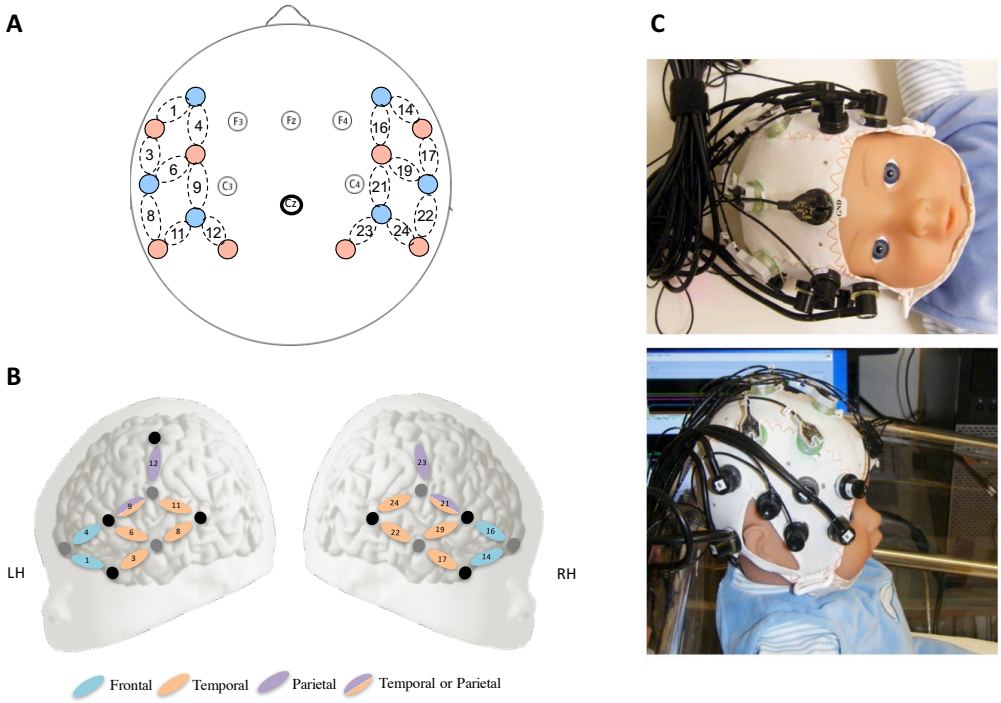


**Fig. 2. Placement of the NIRS optodes and EEG electrodes. A)** Schematic representation of the sources of NIR lights (red circles) and detectors (blue circles). The squares represent the NIRS channels (i.e., coupled sources and detectors). The EEG electrodes are represented by open circles. **B)** Configuration of probe sets overlaid on a schematic newborn brain. For each fNIRS channel located within this probe set, the identity of the underlying brain area (using the LPBA40 atlas) is illustrated according to their localization. The blue channels indicate the position of the probe over the frontal area, the orange channels over the parietal area and the purple over the temporal area on the infant head. Grey circles indicate sources, while black circles indicate detector. **C)** Pictures of the cap on a newborn model doll (Photo credit: Judit Gervain, Laboratoire de Psychologie de la Perception).

Changes in the hemodynamic responses were recorded using NIRS optical probes located on the left and right fronto-temporal regions of the newborns' head (Fig. 2). This localization was based on previous NIRS studies testing speech perception in newborns (5, 26). Two types of analysis were conducted on the recorded hemodynamic activity measured as changes in the concentration of oxygenated (oxyHb) and deoxygenated (deoxyHb)

hemoglobin as a function of auditory stimulation. First, cluster-based permutation analyses using paired samples *t*-tests were conducted to compare concentration changes against a zero baseline in each sound condition. A cluster-based permutation test using a one-way ANOVA was carried out to directly compare the three conditions. The results were followed up with permutation tests using paired samples *t*-tests to identify which pairwise comparisons drove the effects in the permutation test containing the ANOVA. This series of analyses helped to identify the time windows and brain regions of interest (ROI) showing significant activation to auditory stimuli in a data-driven way. Moreover, permutation tests have the advantage of controlling for the multiple comparisons without loss of statistical power, which typically occurs when Bonferroni or other corrections are applied to infant NIRS data (27). Linear mixed effect models were then used to assess the effect of sound conditions (Intact / Fast / Slow), block of stimulation (1 to 6), and ROI derived from the permutation tests on the recorded oxyHb concentration changes. The electrophysiological responses were recorded from EEG electrodes fronto-centrally located on the newborns' head (F3, F4, Cz, C3, C4 according to the 10-20 system, Fig. 2). This localization was based on previous EEG and EEG-NIRS co-recording studies testing speech perception in newborns (28). The amplitude of the EEG responses was averaged in each vocoder condition independently for standard sounds and deviant sounds. A linear mixed effects model was then used in each vocoder condition to assess whether the two types of syllables evoked different EEG responses, known as the mismatch response, reflecting an auditory change detection.

We predicted that newborns' hemodynamic activity should be similar between the intact and the fast conditions, if infants can rely on the fast temporal envelope for phoneme discrimination, as previous studies with 3- to 6-month-olds suggest (20, 21). By contrast, the slow condition may not convey enough fine-grained acoustic details for the newborn brain to process it similarly to the original signal. For phoneme discrimination more specifically, as a

measure of the validity of our design, we expected a significant difference between the standard and deviant syllables in the intact condition, as young infants are known to be able to differentiate syllables differing in a consonant (29). A mismatch response in the other two conditions would indicate that infants can also detect the consonant change using the degraded speech signals.

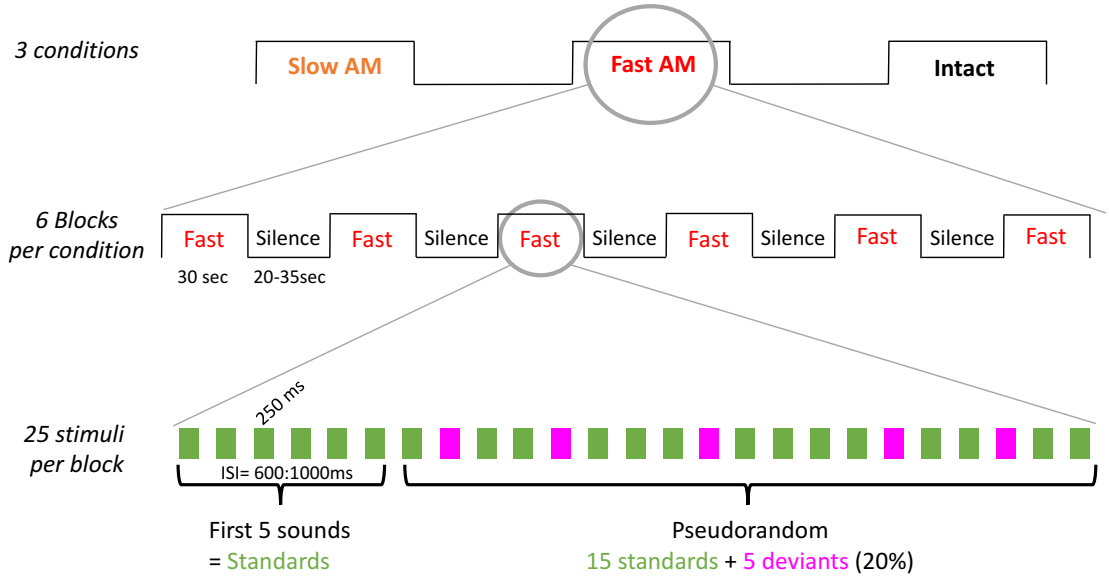


**Fig. 3. Schematic of the experimental design.** Each infant listened to the three sound conditions within 18 consecutive blocks for about 22 min (6 blocks per condition). The Intact condition was always played last and the order of the Slow and Fast conditions were counterbalanced between babies. In each block 25 syllables were played following an odd-ball paradigm, with 20% deviant syllables.

## Results

***How temporal information in speech is processed by the newborn brain: fNIRS results***

The grand average oxyHb NIRS results of the 23 newborns tested are shown in Fig. 4A (for ease of exposition, deoxyHb results are shown separately, Fig. S1). Cluster-based permutation tests (26, 28) comparing each condition to the baseline showed significant

changes in oxyHb concentrations in a time window between 5 and 25 sec after the onset of stimulation. As shown in Fig. 4A, activation was significantly different from the baseline in channel 8 (LH) and 21 (RH) for the Intact condition, in channels 1, 3, 4, 6, 12 (LH) and 14, 17 (RH) for the Fast condition, and in channels 3 (LH) and 16, 17 (RH) for the Slow condition (for all permutation tests, $p < .0001$). For the Fast condition, these significant results indicated a deactivation (negative oxyHb response), whereas for the Intact and the Slow conditions, activation was greater than the baseline. Similar analyses over deoxyHb concentrations showed significant changes in the Intact condition vs. the baseline (in channel 9 between 13 and 25 sec after stimulation onset, and in channel 21 between 10 and 16 sec), and significant changes in the Fast condition vs. the baseline (in channel 11 between 28 and 32 sec). No significant activation compared to the baseline was observed for the Slow condition.

Cluster-based permutation tests were also used to compare changes in oxyHb concentration between conditions. As shown in Fig. 4B, the permutation test comparing all three conditions in a one-way ANOVA yielded significant differences between the conditions in channels 1, 3, 4, 6 and 24 ($p < .01$). Of these, channels 3, 4 and 6 formed a spatial cluster in the LH and channel 24 in the RH ($p < .01$). To follow up on the ANOVA, we conducted permutation tests with paired samples *t*-tests comparing the conditions pairwise. The Intact condition evoked significantly greater activation than the Fast condition in the LH channels 1, 3, 4, and in the RH channels 14, 22, 23 and 24. Of these channels 1 and 3 formed a spatial cluster in the RH ($p = .027$) and channels 22 and 24 in the RH ($p = .046$). Responses in the Slow condition were significantly greater than in the Intact condition in LH channel 1 ($p = 0.039$) and RH channels 22 ($p = .035$). The Slow condition evoked significantly greater activation than the Fast condition in the LH channels 1, 3, 4 and 6 and in the RH channels 14 and 17 ($p < .01$). Of these, channels 1, 3 and 4 in the LH formed a statistically significant

spatial cluster ($p$ = .016), while channels 14 and 17 formed a marginally significant cluster in RH ($p$ = .065). Furthermore, no significant differences in changes in deoxyHb were found between the three conditions (Fig. S1). Consequently, no grand ANOVA was conducted over deoxyHb data.

On the basis of the results of the ANOVA-based permutation test on oxyHb concentration, we selected the fronto-temporal channels 1, 3, 4, and 6 in the LH and 14, 16, 17, and 19 in the RH as ROIs for the Linear Mixed Effects Model. Channels 1, 3, 4 and 6 in the LH were identified as the ROI by the cluster-based permutation test comparing the three conditions, and to have a balanced statistical test, we used the analogous channels in the RH as the ROI for that hemisphere. Linear Mixed Effects Models were then run over average oxyHb concentration changes between 5 and 25 sec after the onset of stimulation, i.e. the time window identified by the cluster-based permutation tests, to assess the effects of Condition (Intact vs. Fast vs. Slow), Hemisphere (Left vs. Right), Channel (4 per hemisphere) and Stimulation Block (1 to 6). Of all the possible models built, the best fitting one included the fixed factors Condition and Block with Participants as a random factor. This model revealed a main effect of Condition [$F(2, 2474) = 10.62$, $p < .001$; Intact vs Fast $p < .001$, Fast vs Slow $p < .001$, Intact vs Slow $p = .62$], Block [$F(5, 2472) = 4.13$, $p < .001$] and a Condition x Block interaction [$F(10, 2473) = 3.69$, $p < .001$]. The main effect of Condition was due to greater responses in the Intact and Slow conditions than in the Fast condition. The main effect of Block reflected a gradual decrease in neural activity in the later blocks as a result of neural habituation often observed in infants' NIRS responses (30). The interaction between Block and Condition revealed that oxyHb concentrations in the Intact and Slow conditions differed in Block 1 and that the activation in Fast and Slow conditions were different in Block 3 as shown in Fig. 5.

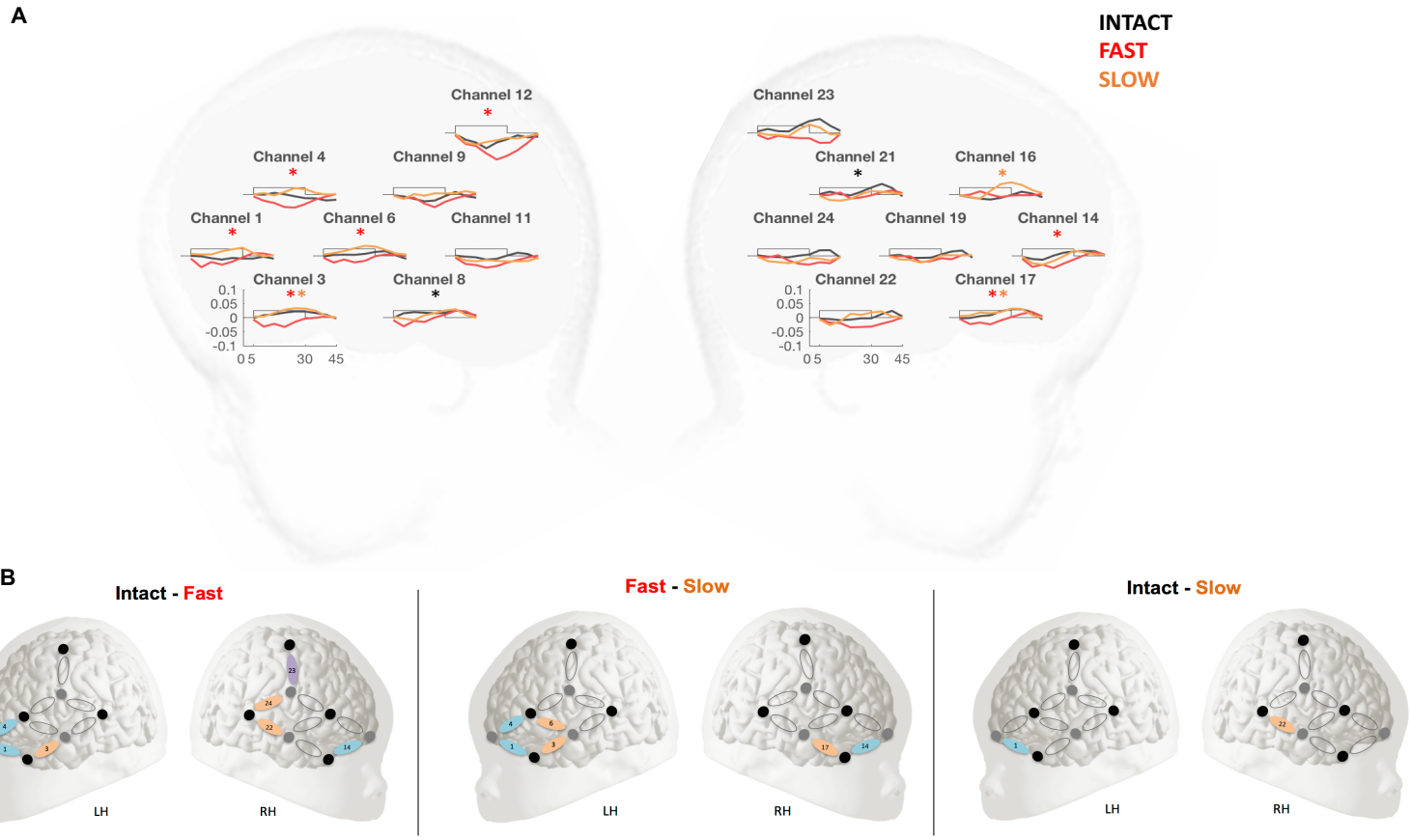


**Fig. 4. Variations in oxyHb. A)** OxyHb concentration changes over stimulation blocks in each channel and in each hemisphere. The x-axis represents time in seconds and the y-axis concentration in mmol-mm. The rectangle along the x-axis indicates time of stimulation. The black lines represent the concentration for the Intact condition, the red for the Fast and the orange for the Slow condition. Color-coded * represents the

channels differing from baseline for each condition ($p < .05$). **B)** Condition-by-condition comparisons of the significantly activated channels according to permutation test ($p < .05$).

In sum, different responses were observed in the three vocoder conditions, with the Slow and Intact conditions evoking positive activation mainly in the left fronto-temporal areas, and the Fast condition yielding a gradual deactivation over time bilaterally.

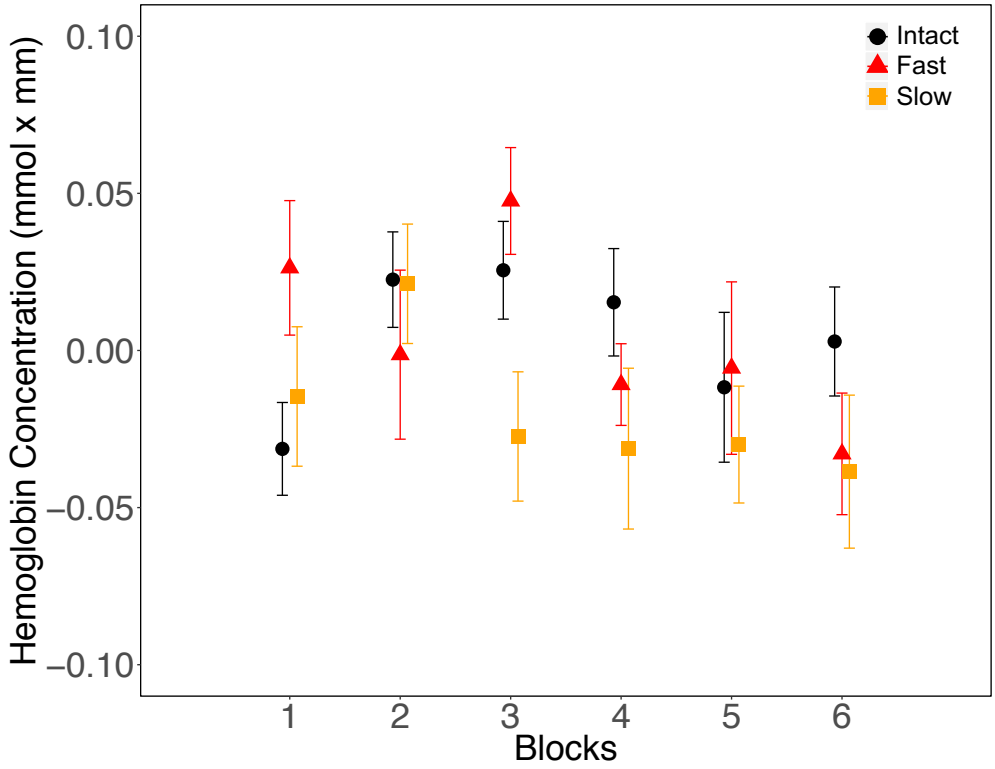


**Fig. 5. Variations in oxyHb over blocks of stimulation.** Averaged oxygenated hemoglobin concentration as a function of stimulation blocks in each vocoder condition in the ROI (average of channels 1, 3, 4, 6, 14, 16, 17 and 19). Error bars represent 1 standard error.

***Phoneme discrimination on the basis of degraded speech signals: EEG results***

Fig. 6 shows the grand average of the EEG responses recorded at F3 for standard and deviant consonants in each condition. We ran linear mixed effects models with fixed factors Trial Type (Standard/Deviant) and Window (8 bins) to assess whether the amplitude of the EEG response recorded for Standard and Deviant sounds was different between 300 and 700 ms after stimulus onset, the usual time window for phoneme discrimination mismatch effects (31). This time window was divided into 8 bins of 50 ms to evaluate the latency of the neural

responses. In each condition, the best fitting model included the fixed factors Trial Type and Window with Participants as a random factor. In all three conditions, a significant main effect of Trial Type was observed indicating that deviant and standard consonants elicited different activations in each sound condition [Intact: $F(1, 270) = 5.22$, $p = .023$, $\eta^2 = .019$; Fast: $F(1, 270) = 16.69$, $p < .001$, $\eta^2 = .058$; Slow: $F(1, 270) = 8.37$, $p = .004$, $\eta^2 = .03$]. The mismatch response was positive in the Intact condition, but negative in the Fast and Slow conditions. A main effect of Window was also observed in the Intact and Slow conditions [$F(7, 270) = 2.238$, $p = .032$, $\eta^2 = .055$; $F(7, 270) = 2.483$, $p = .017$, $\eta^2 = .06$, respectively] due to more positive responses in the latter time windows (500-700 ms) compared to the first ones (300-450 ms). This effect was marginal in the Fast condition [$F(7, 270) = 2.004$, $p = .055$, $\eta^2 = .049$]. No Trial Type x Window interaction was observed in any condition [Intact: $F(7,270) = .313$, $p = .948$; Fast: $F(7,270) = .811$, $p = .578$; Slow: $F(7,270) = .079$, $p = .99$]. Thus, in each condition, responses to the Deviant differed from the Standard starting from 300 to 700 ms after stimulus onset.

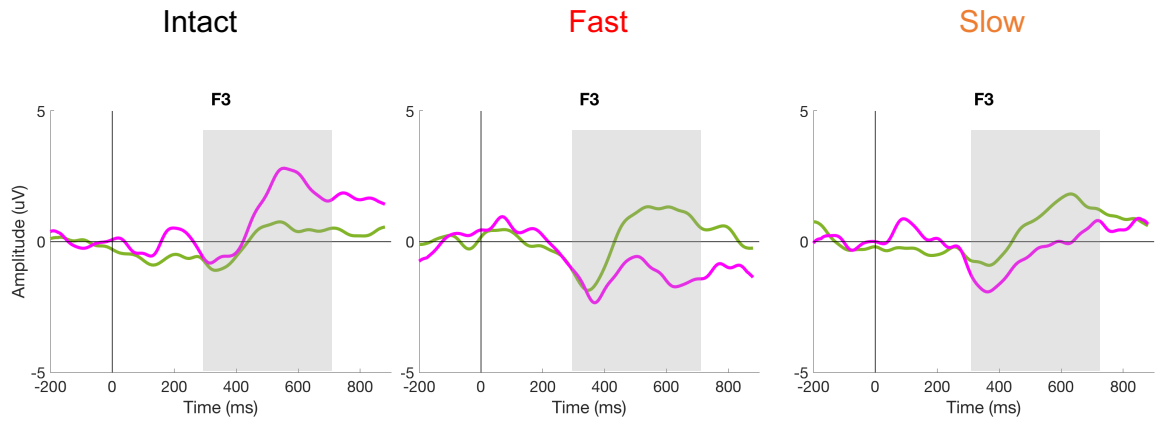


**Fig. 6. Grand average of EEG responses.** Group mean amplitude variations (μV) of EEG responses recorded at F3 for the Standard (green lines) and the Deviant (magenta lines) in each condition for the group of newborns exposed to a consonant change. Responses to Standard and Deviant differ from each other in each condition in the time window 300-700 ms represented by the grey rectangle.

## Discussion

The present NIRS-EEG co-registration study indicates that like adults, neonates do not require the temporal fine structure and fast envelope information to distinguish consonant features in quiet. However, the fast and slow components of the speech envelope are processed differently at birth.

First, the electrophysiological results demonstrate that the neonate brain is able to detect a change in place of articulation between two French stop consonants as shown by the mismatch EEG response in the Intact condition. Consistently with some previous studies with young infants, we observed a positive mismatch response (32). Newborns were also able to detect the consonant change on the basis of the envelope cues without the temporal fine structure as well as on the basis of the slow temporal variations alone (AM < 8 Hz). These results are consistent with behavioral data obtained with older infants and adults (13, 14, 20, 21), for whom the slowest envelope cues are also sufficient to detect consonant changes in silence. However, the direction of the mismatch response differed between the Intact and the degraded conditions. Different polarities for deviant sounds have been observed in previous studies according to difference in the design, inter-stimulus interval or reference electrode (33). These methodological factors cannot explain our results, as the polarity difference occurred within the same study. Polarity reversals were also observed in infants as a function of task difficulty, e.g. for more challenging acoustic discriminations, for instance for small pitch differences (34). This implies that the difficulty of change detection may have been different in the Intact as compared to the two degraded conditions. This hypothesis needs to be further investigated at different ages to track the role of temporal modulations in speech perception during early development.

However, it is important to note that, while discrimination was possible in all three conditions, as indicated by brain potentials, the full and slow envelope cues were not

processed in the same way. The hemodynamic responses differed in their magnitude, time course and localization according to the vocoder condition, that is, as a function of the temporal information available in the speech signal. More precisely, the NIRS recordings revealed bilateral activations for the Slow condition and a more left-lateralized one for the Fast condition. This different pattern of activation may reflect adult-like brain specialization for the slow and fast modulation cues already at birth. Previous studies in adults suggest that slow modulations preferentially activate the right hemisphere, while the faster temporal modulations preferentially activate the left hemisphere (8, 9, 35, 36), although some recent models have called this simple division of labor view into question (10). In adults, the fast and slow rates of the temporal envelope are processed by both distinct and shared neural substrates. One neuroimaging study using vocoded-speech sounds showed similar activation for original speech and noise-vocoded speech preserving slow and fast AM (< 320 Hz, extracted in 6 channels) in the superior temporal gyrus, but dissimilar activations in the superior temporal sulcus (37), demonstrating that the processing of the full speech signal and of fast AM cues do not involve the same neural processing in adults. Slow and fast modulations have not been directly compared in adults using speech sounds. For non-speech sounds, e.g. white noise, fMRI studies found different brain responses between AM rates below 16 Hz and above 128 Hz (8). More specifically, slow and fast envelope rates in non-speech sounds activated the same cortical regions (superior temporal gyrus and sulcus), but the time courses of the activation differed according to the AM rate. Interestingly, responses recorded bilaterally in Heschl's gyrus were tuned to the slowest AM rates, 8 Hz, and the faster modulations of non-speech sounds have been shown to activate preferentially the LH (9). The present findings provide unique insight showing that the newborn brain also exhibits differential processing for different AM rates in speech, and suggest that a differential hemispheric specialization for the processing of slow and fast envelope information is

already present at birth. The human brain is thus already tuned to fast and slow AM information in speech from the get-go, laying the foundations of later language learning and speech comprehension.

It is noteworthy that the time course of the hemodynamic responses was more similar between the Intact and Slow conditions than between the Intact and the Fast conditions. This is surprising given that the slow condition is more acoustically degraded, i.e. less similar to the Intact condition, than the Fast one. Deactivation (negative oxyHb response) is often observed in newborn studies (30) and may be related to neural habituation due to stimulus repetition. In the present study, it is possible that the sharply decreasing hemodynamic response over time in the Fast condition may reflect faster neural habituation in this condition than in the other two conditions. This deactivation cannot be due to systemic variations in blood flow, as the fast and slow conditions were presented in a counterbalanced order. This result is consistent with previous studies using non-speech sounds (22) showing a specific neural response for relatively fast temporal modulations (change every 25 ms that is equivalent to ~ 40 Hz fluctuations), but not for slower modulations in neonates. Moreover, a recent MEG study showed that fetuses are able to detect slow and fast AM rates (from 2 to 91 Hz) modulating non-speech sounds and that they show the highest response to 27 Hz modulations, assumed to be better transmitted by bone conduction than slower AM rates (38). Responses to 4 Hz were progressively maturing over the 31$^{st}$ and 39$^{th}$ gestational week. The development of these specific auditory responses has not been clearly related to the development of speech perception yet. One may hypothesize that the present responses for relatively fast temporal modulations of speech observed at birth might reflect that infants depend more heavily on fast-envelope cues than any other modulation cues. Because fast envelope cues carry more information about fundamental frequency and formant transitions that slow envelope cues, this specific response may be consistent with infants' early ability to

detect phonetic difference and their preference for exaggerated prosodic cues (39). While we cannot be certain that the present vocoded syllables were processed as speech, there are two pieces of evidence pointing in this direction. First, there is overlap in the localization of the activated channels in the three conditions, including the intact condition, which is undeniably speech. Second, these activated channels are in the temporal and inferior frontal areas, i.e. in the auditory and language network. The fact that activation was bilateral does not argue against the sounds being processed as speech, since bilateral activation in response to speech is commonly observed in newborns (40, 41).

Future studies with young infants are needed to fully characterize the maturation of the auditory pathway in order to better understand the interplay between auditory development and language development. During the first year of life, infants become better at discriminating the speech contrasts of their native language, but do not show the same improvement for non-native contrasts (42). This phenomenon is called perceptual attunement, or speech specialization. It is possible that with a given auditory experience, listeners learn to ignore specific acoustic cues of speech irrelevant to develop native-language categories (20). Thus, the reliance on fast speech AM cues may change with greater exposure to the native speech sounds. More studies comparing the reliance upon the acoustic cues of the speech signal during early development are needed to explore this hypothesis. The advantage of the psychoacoustic approach for psycholinguistic studies is to describe the role of fine spectral and temporal modulations, which have neuro-correlates in the auditory system, for speech perception. Thus, this approach offers a new opportunity to characterize precisely the auditory sensory mechanisms involved in speech processing during early development, that is during a critical period for language development.

In sum, our study demonstrates for the first time that the human auditory system is able to encode speech in fine details on the basis of highly reduced acoustic information,

specifically the slowest amplitude modulation cues, already from birth. Furthermore, the newborn brain already shows considerable specialization to different temporal cues in the speech signal, laying the foundations of infants' astonishingly sophisticated speech perception and language learning abilities.

## Material and Methods

*Participants.* Newborn infants born at a gestational age between 37 and 42 weeks, with Apgar scores ≥ 8 in the 1st and 5th minutes following birth, a head circumference greater than 32 cm, and having no known neurological or hearing abnormalities were recruited for the study at the maternity ward of the Robert Debré Hospital, Paris, France. Newborns' hearing was assessed by a measurement of auditory brainstem responses during their stay at the hospital. The study was approved by the research ethics committee of University Paris Descartes (CERES approval nr. 2011-13) and all parents provided written informed consent prior to participation.

A group of 74 healthy full-term neonates (mean age: 1.8 days, range: 1-4 days) were tested. Thirteen newborns did not complete the study due to crying (n=10) and parental/external interference (n=3). Of the 61 infants who completed the study, 7 were excluded from the NIRS analysis because of technical problems during the recording and 31 due to poor data quality (large motion artifacts or noise). A total of 23 newborns were included in the fNIRS analyses (16 females). Of the 61 completers, 3 were excluded from the EEG analysis due to technical problems and 38 because of poor data quality (artifacts). A total of 20 newborns were thus included in the EEG analyses (11 females). The mothers of all infants spoke French during the pregnancy, but 10 also spoke a second language around 50% of the time (Arabic, Bambara, Italian, Kabyle, Mandarin, Portuguese, Soninke or Swahili).

*Stimuli.* Eight natural occurrences of the syllables /pa/ and /ta/ were recorded by a native French speaker, who was instructed to speak clearly. All tokens were comparable in duration (mean = 194 ms, SD = 14 ms) and F0 (213 Hz, SD = 4 Hz). All stimuli were equated in global root-mean-square (RMS) level. Each stimulus was processed in three vocoder conditions. Specifically, three different tone-excited vocoders were designed. In each condition, the original speech signal was passed through a bank of 32 2nd-order gammatone filters (43), each 1-equivalent rectangular bandwidth (ERB) wide with center frequencies (CFs) uniformly spaced along an ERB scale ranging from 80 to 8,020 Hz. A Hilbert transform was then applied to each bandpass filtered speech signal to extract the envelope component and temporal fine structure carrier. The envelope component was low-pass filtered using a zero-phase Butterworth filter (36 dB/octave rolloff) with a cutoff frequency set to either $ERB_N/2$ (Intact and Fast condition) or 8 Hz (Slow condition). In the Fast and Slow conditions, the temporal fine structure carrier in each frequency band was replaced by a sine wave carrier with a frequency corresponding to the center frequency of the gammatone filter and a random starting phase. Each tone carrier was then multiplied by the corresponding filtered envelope function. In the Intact condition, the original temporal fine structure was multiplied by the filtered envelope function in each band. The narrow-band speech signals were then added up and the level of the wideband speech signal was adjusted to have the same RMS value as the input signal in each condition. Thus, in the Intact condition, the resulting speech signal contained the original envelope and original temporal fine structure in 32 bands. In the Fast condition, the vocoder manipulation discarded the original (within channel) temporal fine structure cues, but retained the fast envelope cues (cutoff frequency set to $ERB_N/2$). In the Slow condition, the manipulation discarded both the original temporal fine structure and the fast envelope cues to preserve only the slowest envelope information in each band (< 8Hz). Syllabic information was thus preserved in both degraded conditions, but

voice-pitch and formant transition information was preserved only in the Fast condition, and drastically reduced in the Slow condition.

*Equipment and procedure.* Optical imaging was performed with a NIRScout 816 machine (NIRx Medizintechnik GmbH, Berlin, Germany), using pulsated LED sequential illumination with two wavelengths of 760 nm and 850 nm to record the NIRS signal at a 15.625 Hz sampling rate. Three LED sources were placed on each side of the head in analogous positions, and were illuminated sequentially. They were coupled with 4 detectors on each side of the head. The configuration of the 16 channels (8 per hemisphere) created with the 3 sources and 4 detectors per hemisphere is shown in Fig. 2. We embedded the optodes in an elastic cap (EasyCap). The source-detector separation was 3 cm. For each infant, we selected a cap size according to their head circumference. We also adjusted the cap according to Cz measurement and ear location. Localization analysis for our newborn headgear was performed as in (26). The electrophysiological recording was performed with a BrainProducts actiCHamp EEG amplifier (BrainProducts GmbH, Munich, Germany) and active electrodes. Five active electrodes (F3, Fz, F4, C3, C4, 10-20-system), embedded in the same cap as the NIRS optodes, were used to record the EEG signal at a 2000 Hz sampling rate, referenced to the vertex (Cz). The stimuli were played through two speakers elevated to the height of the crib, approximately 30 cm from the infants' head on each side and at around 70 dB SPL.

While newborns were lying quietly in their hospital cribs, the syllables were presented to them in long stimulation blocks (30 sec) with 6 blocks per vocoder condition. The inter-stimulus interval between syllables within a block was varied randomly between 600 and 1000 ms and the inter-block interval between 20 and 35 sec. The Intact sound condition was always played last while the order of the Slow and Fast conditions was counterbalanced

across babies. Each block contained 25 syllables, out of which 20 were standard syllables (e.g. /pa/) and 5 were deviants (e.g. /ta/), allowing an event-related assessment of the responses to individual syllables within blocks similarly to the classical oddball or mismatch design in EEG studies. Thus, each block of stimulation comprised a ratio of 80-20% of standard and deviant sounds. The first 5 sounds were always standards to allow expectations about the standard to build up. The standard and deviant syllables were counterbalanced across babies. The whole experiment lasted around 22 min.

*Data analysis. fNIRS.* Analyses were conducted on oxyHb and deoxyHb. Data were band-pass filtered between 0.01 and 0.7 Hz to remove low-frequency noise (i.e., slow drifts in Hb concentrations) as well as high frequency noise (i.e., heartbeat). Movement artifacts were removed by identifying block-channel pairs in which a change in concentration greater than 0.1 mmol × mm over a period of 0.2 s, occurred, and rejecting the block for that channel. Channels with valid data for less than 3 out of 6 blocks per condition were discarded. A baseline was established by using a linear fit over the 5s time window preceding the onset of the block and the 5s window beginning 15s after the end of the block. The 15 s resting period after stimulus offset was used to allow the hemodynamic response function (HRF) to return to baseline. Analyses were conducted in MATLAB (version R2015b) with custom analysis scripts.

Regions of interest (ROI) were defined according to the permutation analyses. For the cluster-based permutation test, we used paired-sample *t*-tests to compare each vocoder condition to a zero baseline in each channel. Then all temporally and spatially adjacent pairs with a t-value greater than a standard threshold (we used $t = 2$) were grouped together into cluster candidates. We calculated cluster-level statistics for each cluster candidate by summing the *t*-values from the *t*-tests for every data point included in the cluster candidate.

We then identified the cluster candidate with the largest *t*-value for each hemisphere. Then a permutation analysis evaluated whether this cluster level statistic was significantly different from chance. This was done by randomly labelling the data as belonging to one or the other experimental condition. The same *t*-test statistic as before was computed for each random assignment, which allowed us to obtain its empirical distribution under the null hypothesis of no difference between the baseline and each condition. Clusters were then formed as before. The proportion of random partitions that produce a cluster-level statistic greater than the actually observed one provides the *p* value of the test. In all, 100 permutations under the null hypothesis were conducted for this robust comparison.

Similar permutation tests were run to directly compare the three conditions, except that the *t*-test was replaced with a one-way ANOVA with factor Condition (Intact vs. Fast vs. Slow) and 1000 permutations were conducted.

The channels identified by the permutation tests as spatial clusters were included in Linear Mixed Effect Models comparing the effects of the fixed factors Condition (Intact vs. Fast vs. Slow), Block of stimulation (1-6), Channel and Hemisphere (LH vs. RH) and the random effect of Participant on the variations in oxy-Hb concentration. The best fitting model is reported and interpreted.

*EEG.* The EEG signal was re-sampled at 200 Hz and band-pass filtered at 0.5 to 20 Hz (31, 32). The continuous EEG data was segmented into epochs of 1000 ms including a 200 ms baseline (-200-0msec) and time-locked to stimulus onset. Epochs including the first standard sound of each vocoder condition, and standards directly following deviant sounds were excluded to avoid large dishabituation/novelty detection responses. Epochs with abnormal values (< -120 μV and > +120 μV) were then excluded automatically. Infants were retained for data analysis if the number of deviant trials was at least 10 in each condition. For the

group of infants included in the final analysis, the average number of deviant epochs retained for analysis was 28 in the Intact condition and 27 in both the Fast and Slow conditions.

For statistical purposes, we averaged together all good deviant epochs and all good standard epochs in each condition for each infant in each channel. Based on visual inspection (see Supplementary Information Fig. S2), the EEG amplitude recorded at F3 was averaged between 300 and 700 ms after stimulus onset to assess the mismatch response in each condition.

**Acknowledgments:**

The authors wish to thank all the parents and infants who participated in the study and all the personnel of Hopital Robert Debre, Paris, France. This work was supported by an Emergence(s) Programme Grant from the City of Paris, a Human Frontiers Science Program Young Investigator Grant (RGY-0073-2014) as well as the ERC Consolidator Grant "BabyRhythm" (nr. 773202) to JG. LC is currently supported by an ANR grant nr-17-CE28-008.

LC and JG designed the experiment ; LC performed the research and analyzed the data, LC and JG wrote the paper.

Competing Interests: The authors declare that they have no competing interests

Data Sharing Statement. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data available from authors upon request.

## References


1. J. K. Moore, Maturation of human auditory cortex: implications for speech perception. *Ann Orni Hhim Uiryngoi* **11**, 2 (2002).

2. C. Abdala, A longitudinal study of distortion product otoacoustic emission ipsilateral suppression and input/output characteristics in human neonates. *J. Acoust. Soc. Am.* **114**, 3239–3250 (2003).

3. P. K. Kuhl, Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* **5**, 831–843 (2004).

4. 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).

5. J. Gervain, F. Macagno, S. Cogoi, M. Peña, J. Mehler, The neonate brain detects speech structure. *Proc. Natl. Acad. Sci.* **105**, 14222–14227 (2008).

6. C. Moon, R. P. Cooper, W. P. Fifer, Two-day-olds prefer their native language. *Infant Behav. Dev.* **16**, 495–500 (1993).

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

8. A. L. Giraud, *et al.*, Representation of the temporal envelope of sounds in the human brain. *J. Neurophysiol.* **84**, 1588–1598 (2000).

9. C. Liégeois-Chauvel, C. Lorenzi, A. Trébuchon, J. Régis, P. Chauvel, Temporal envelope processing in the human left and right auditory cortices. *Cereb. Cortex* **14**, 731–740 (2004).

10. D. Poeppel, The neuroanatomic and neurophysiological infrastructure for speech and language. *Curr. Opin. Neurobiol.* **28**, 142–149 (2014).

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

12. R. J. Zatorre, P. Belin, Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* **11**, 946–953 (2001).

13. R. Drullman, Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am.* **97**, 585–592 (1995).

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

15. L. Xu, B. E. Pfingst, Relative importance of temporal envelope and fine structure in lexical-tone perception. *J. Acoust. Soc. Am.* **114**, 3024–3027 (2003).

16. F.-G. Zeng, *et al.*, Speech recognition with amplitude and frequency modulations. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2293–2298 (2005).

17. J. Bertoncini, T. Nazzi, L. Cabrera, C. Lorenzi, Six-month-old infants discriminate voicing on the basis of temporal envelope cues. *J. Acoust. Soc. Am.* **129**, 2761–2764 (2011).

18. L. Cabrera, J. Bertoncini, C. Lorenzi, Perception of Speech Modulation Cues by 6-Month-Old Infants. *J. Speech Lang. Hear. Res.* **56**, 1733–1744 (2013).

19. A. D. Warner-Czyz, D. M. Houston, L. S. Hynan, Vowel discrimination by hearing infants as a function of number of spectral channels a. *J. Acoust. Soc. Am.* **135**, 3017–3024 (2014).

20. L. Cabrera, C. Lorenzi, J. Bertoncini, Infants discriminate voicing and place of articulation with reduced spectral and temporal modulation cues. *J. Speech Lang. Hear. Res.* **58**, 1033–1042 (2015).

21. L. Cabrera, L. Werner, Infants' and Adults' Use of Temporal Cues in Consonant Discrimination. *Ear Hear.* (2017) (April 10, 2017).

22. S. Telkemeyer, *et al.*, Sensitivity of newborn auditory cortex to the temporal structure of sounds. *J. Neurosci. Off. J. Soc. Neurosci.* **29**, 14726–14733 (2009).

23. L. Varnet, M. C. Ortiz-Barajas, R. G. Erra, J. Gervain, C. Lorenzi, A cross-linguistic study of speech modulation spectra. *J. Acoust. Soc. Am.* **142**, 1976–1989 (2017).

24. M. Mahmoudzadeh, *et al.*, Syllabic discrimination in premature human infants prior to complete formation of cortical layers. *Proc. Natl. Acad. Sci.* **110**, 4846–4851 (2013).

25. J. L. Mueller, A. D. Friederici, C. Männel, Auditory perception at the root of language learning. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 15953–15958 (2012).

26. N. Abboub, T. Nazzi, J. Gervain, Prosodic grouping at birth. *Brain Lang.* **162**, 46–59 (2016).

27. T. E. Nichols, A. P. Holmes, Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).

28. S. Benavides-Varela, J. Gervain, Learning word order at birth: A NIRS study. *Dev. Cogn. Neurosci.* **25**, 198–208 (2017).

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

30. C. Bouchon, T. Nazzi, J. Gervain, Hemispheric asymmetries in repetition enhancement and suppression effects in the newborn brain. *PloS One* **10**, e0140160 (2015).

31. G. Dehaene-Lambertz, Cerebral specialization for speech and non-speech stimuli in infants. *J. Cogn. Neurosci.* **12**, 449–460 (2000).

32. G. Dehaene-Lambertz, M. Pena, Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport* **12**, 3155–3158 (2001).

33. O. Martynova, J. Kirjavainen, M. Cheour, Mismatch negativity and late discriminative negativity in sleeping human newborns. *Neurosci. Lett.* **340**, 75–78 (2003).

34. M. L. Morr, V. L. Shafer, J. A. Kreuzer, D. Kurtzberg, Maturation of mismatch negativity in typically developing infants and preschool children. *Ear Hear.* **23**, 118–136 (2002).

35. R. J. Zatorre, P. Belin, V. B. Penhune, Structure and function of auditory cortex: music and speech. *Trends Cogn. Sci.* **6**, 37–46 (2002).

36. D. Poeppel, Pure word deafness and the bilateral processing of the speech code. *Cogn. Sci.* **25**, 679–693 (2001).

37. S. K. Scott, S. Rosen, H. Lang, R. J. S. Wise, Neural correlates of intelligibility in speech investigated with noise vocoded speech—A positron emission tomography study. *J. Acoust. Soc. Am.* **120**, 1075–1083 (2006).

38. R. Draganova, *et al.*, Fetal auditory evoked responses to onset of amplitude modulated sounds. A fetal magnetoencephalography (fMEG) study. *Hear. Res.* **363**, 70–77 (2018).

39. J. Mehler, *et al.*, A precursor of language acquisition in young infants. *Cognition* **29**, 143–178 (1988).

40. H. Sato, *et al.*, Cerebral hemodynamics in newborn infants exposed to speech sounds: A whole-head optical topography study. *Hum. Brain Mapp.* **33**, 2092–2103 (2012).

41. L. May, K. Byers-Heinlein, J. Gervain, J. F. Werker, Language and the newborn brain: does prenatal language experience shape the neonate neural response to speech? *Front. Psychol.* **2**, 222 (2011).

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

43. D. Gnansia, V. Péan, B. Meyer, C. Lorenzi, Effects of spectral smearing and temporal fine structure degradation on speech masking release. *J. Acoust. Soc. Am.* **125**, 4023–4033 (2009).

**Supplementary information: Figure S1 and S2**

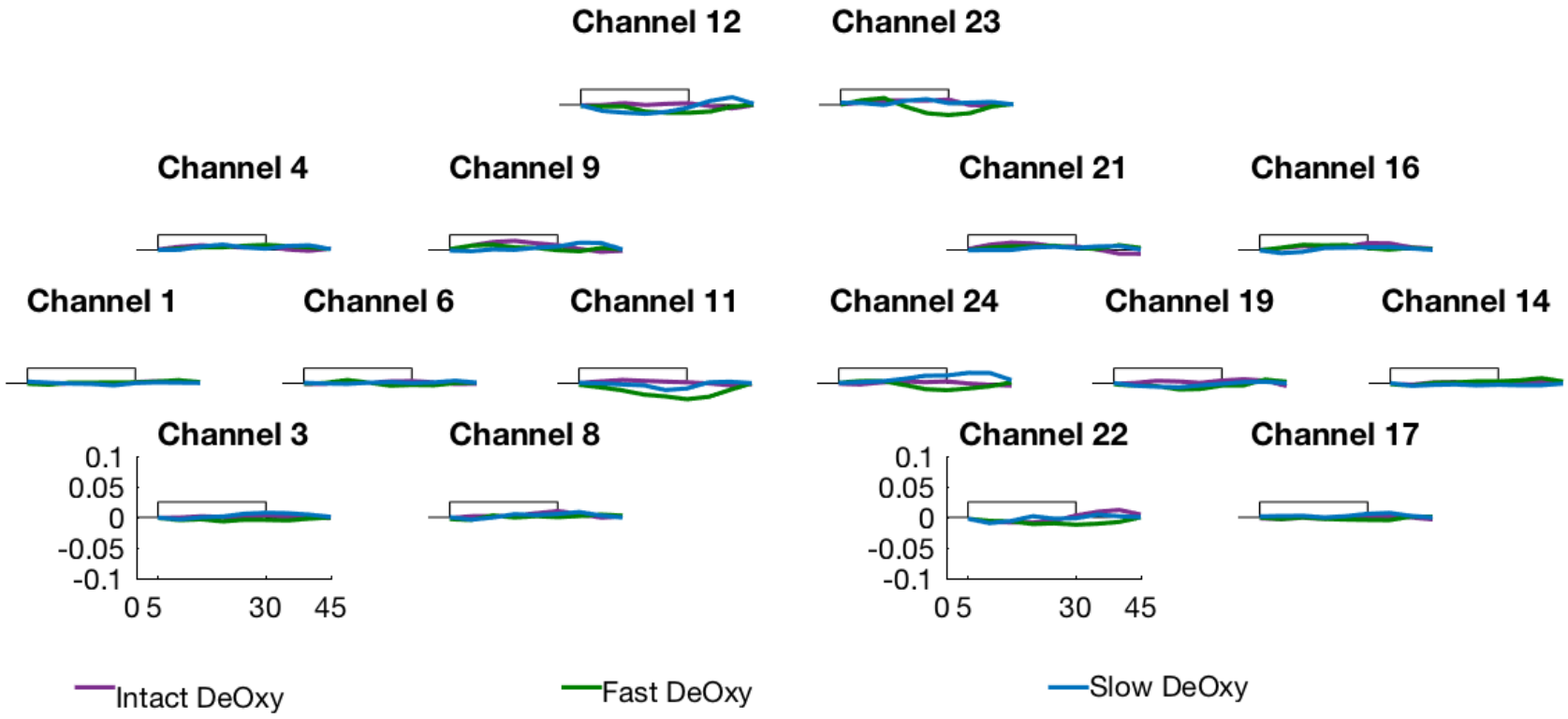


**Fig. S1.** Deoxygenated hemoglobin concentration changes over stimulation blocks in each channel and in each hemisphere. The x-axis represents time in seconds and the y-axis concentration in mmol-mm. The rectangle along the x-axis indicates time of stimulation. The purple lines represent the concentration for the Intact condition, the green for the Fast and the bleu for the Slow condition. Changes in concentration did not differ from baseline in any channel for any condition ($p > .05$).

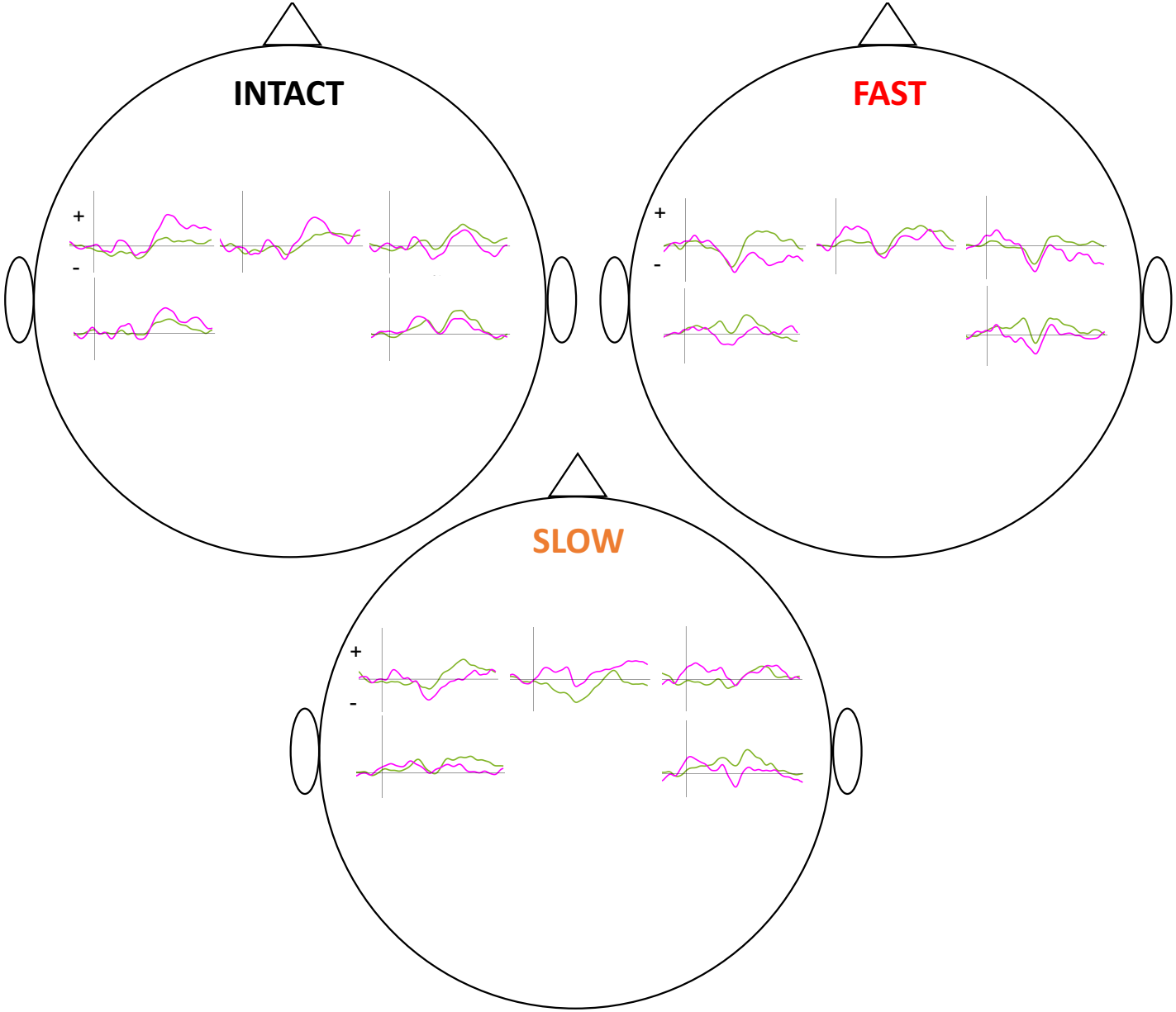


**Fig. S2.** Grand average of amplitude variations (on the y-axis in μV) of EEG responses over time (on the x-axis from – 200 to +800 ms) recorded at F3, Fz, F4 (from left to right upper row, respectively), C3 and C4 (from left to right lower row, respectively) for the Standard (green lines) and the Deviant (magenta lines) in each condition for the group of newborns.