



HAL
open science

Emergence of the cortical encoding of phonetic features in the first year of life

Giovanni M Di Liberto, Adam Attaheri, Giorgia Cantisani, Richard B Reilly,
Áine Ní Choisdealbha, Sinead Rocha, Perrine Brusini, Usha Goswami

► **To cite this version:**

Giovanni M Di Liberto, Adam Attaheri, Giorgia Cantisani, Richard B Reilly, Áine Ní Choisdealbha, et al.. Emergence of the cortical encoding of phonetic features in the first year of life. 2023. hal-04022481

HAL Id: hal-04022481

<https://hal.science/hal-04022481>

Preprint submitted on 10 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Emergence of the cortical encoding of phonetic features in the first year of life

Giovanni M. Di Liberto^{1,2,*}, Adam Attaheri^{2,*}, Giorgia Cantisani^{3,1}, Richard B. Reilly^{4,5}, Áine Ní Choisdealbha², Sinead Rocha², Perrine Brusini², Usha Goswami²

1 ADAPT Centre, School of Computer Science and Statistics, Trinity College, The University of Dublin, Ireland; Trinity College Institute of Neuroscience

2 Centre for Neuroscience in Education, Department of Psychology, University of Cambridge, United Kingdom

3 Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, 75005 Paris, France

4 School of Engineering, Trinity Centre for Biomedical Engineering, Trinity College, The University of Dublin. Trinity College Institute of Neuroscience

5 School of Medicine, Trinity College, The University of Dublin, Ireland

** These authors contributed equally*

Correspondence: diliberg@tcd.ie

Conflicts of interest: none declared.

Funding sources: This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 694786) (A.A., U.G.). This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Trinity College Dublin (G.D.L., G.C.). ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme. This work was also supported by the Science Foundation Ireland Career Development Award 15/CDA/3316 (G.D.L., R.R.). G.C. was supported by an Advanced European Research Council grant (NEUME, 787836) and by the FrontCog grant ANR-17-EURE-0017.

Acknowledgements: We thank Dimitris Panayiotou, Alessia Philips, Natasha Mead, Helen Olawole-Scott, Panagiotis Boutris, Samuel Gibbon, Isabel Williams, Sheila Flanagan, and Christina Grey who helped collecting the data as well as all the families of the infant participants. We thank Dr. Susan Richards for her assistance on the phoneme transcription. We thank the CogHear workshop organisers (Mounya Elhilali, Malcolm Slaney, and Shihab Shamma) and participants for their useful feedback on the early results of this study.

Word count abstract: 148.

Word count (excluding abstract, title page, references and methods): 4114.

34 **Abstract**

35 Even prior to producing their first words, infants are developing a sophisticated speech processing system,
36 with robust word recognition present by 4-6 months of age. These emergent linguistic skills, observed with
37 behavioural investigations, are likely to rely on increasingly sophisticated neural underpinnings. The infant
38 brain is known to robustly track the speech envelope, however to date no cortical tracking study could
39 investigate the emergence of phonetic feature encoding. Here we utilise temporal response functions
40 computed from electrophysiological responses to nursery rhymes to investigate the cortical encoding of
41 phonetic features in a longitudinal cohort of infants when aged 4, 7 and 11 months, as well as adults. The
42 analyses reveal an increasingly detailed and acoustically-invariant phonetic encoding over the first year of
43 life, providing the first direct evidence that the pre-verbal human cortex learns phonetic categories. By 11
44 months of age, however, infants still did not exhibit adult-like encoding.

45

46 The human ability to understand speech relies on a complex neural system, whose foundations develop over
47 the first few years of life. A wealth of evidence on the developmental progression of speech perception is
48 available from infant behavioural studies, including with neonates, augmented by studies of speech
49 production from around the second year of life^{1,2}. Yet our understanding of speech perception in the first
50 year of life is largely dependent on tasks relying on simple behaviours (e.g., head turn preference procedure).
51 Direct investigation of the neural encoding of phonetic information in continuous natural speech across the
52 first year of life has not previously been possible. Experiments using behavioural measures enable the
53 assessment of valuable factors such as the familiarity of a particular speaker, the phonetic features that can
54 be discriminated, and sensitivity to native versus non-native speech contrasts, thereby providing a time-line
55 for the development of speech perception in the first year of life¹. However, behavioural methods can only
56 serve as an indirect index of the emergence of linguistic skills, and cannot reveal when the phonetic encoding
57 in the human cortex becomes invariant across different instantiations. Previous behavioural studies focused
58 on sound discrimination due to methodological constraints, and made use of targeted experimental
59 paradigms involving simple stimuli. Although this behavioural timeline has been complemented by
60 neurophysiological investigations, these studies have employed similar targeted paradigms, with the most
61 widely-used neurophysiological measure with infants being the mismatch negativity (MMN, or mismatch
62 response, MMR). The MMR is a neurophysiological signature of automatic change detection³⁻⁵ typically used
63 to measure the ability to discriminate particular speech contrasts. However, previous studies showed that
64 such mismatch responses in infants can sometimes be positive⁶, causing inconsistencies that can complicate
65 or limit their use in infants. This leaves us with a number of key open questions: 1) How do infants perceive
66 and encode the phonological units such as syllables and phonemes in **continuous natural speech**? 2) How
67 are these speech sounds **encoded** in the infant brain? And 3) how does that encoding **develop across the**
68 **first year of life?**

69 This study is the first to address these research questions directly. Non-invasive electroencephalography
70 signals (EEG) were recorded as infants listened to 18 nursery rhymes (vocals only with no instruments
71 involved) through video recordings of a native English speaker. EEG recordings were carried out at 4, 7 and

72 11 months of age from the first 50 participants in a longitudinal cohort involving 122 infants (the same
73 subjects were tested in the three subsequent sessions and only participants with all sessions were selected).
74 Three participants were excluded due to excessive EEG noise (see **Methods**). We then measured how the
75 infant brain encodes acoustic and phonetic information by means of the multivariate Temporal Response
76 Function analysis (TRF), a neurophysiology framework enabling the study of how neural signals encode
77 continuous sensory stimuli^{7,8}. TRF analyses were also carried out on recordings from adult participants
78 listening to the same stimuli. We targeted one key aspect for speech perception, the perception of phonetic
79 features. We do not assume here that encoding phonetic features equates to encoding phonemes, as there
80 is a large psychoacoustic and developmental literature showing that phonemes are only represented by
81 literate brains^{9,10}. Our core hypothesis was rather that phonetic feature encoding (invariant to acoustic
82 changes) would emerge in the neural responses to natural speech during the first year of life.

83 Speech TRFs reflect the neural tracking of (or neural entrainment to, in the broad sense¹¹) natural speech
84 features (e.g., acoustic envelope), offering a direct window into human perception during natural listening
85 without imposing any particular task other than listening. In recent years, neural tracking measures have
86 played a growing role in the study of speech comprehension and auditory processing in general. Many TRF
87 studies have assessed the neural tracking of the acoustic envelope¹²⁻¹⁵, which is an important property of
88 speech that co-varies with a number of key properties of interest (e.g., syllable stress patterns, syllables,
89 phonemes). Neural tracking of the speech envelope (or envelope tracking) was shown to reflect both bottom-
90 up and top-down cortical processes in adult listeners, encompassing fundamental functions such as selective
91 attention¹⁵⁻¹⁷, working memory processing load¹⁸, and prediction^{19,20}. While robust envelope tracking has also
92 been demonstrated in infants²¹⁻²⁶, envelope measures only reveal some of the cortical mechanisms
93 underlying speech perception. Recent work with adults and children has demonstrated that TRFs can be
94 extended to isolate the neural encoding of targeted speech properties of interest, starting from phonetic
95 features²⁷. Phonetic encoding was measured in multiple studies from different research teams²⁷⁻³¹, and the
96 neural tracking was shown to correlate with phonemic awareness skills in school-aged children between 6
97 and 12 years of age³² and with second language proficiency in adults³³.

98 Here, we employed TRFs to test the hypothesis that the neural encoding of phonetic features during natural
99 speech listening is already developing during the first-year of life. Current behavioural data indicate that
100 infant perception becomes more selective towards native than non-native speech contrasts around 9-12
101 months of age³⁴ (see footnoteⁱ), with perceptual “magnet” effects helping to isolate native from non-native
102 phonetic contrasts already by 6 months³⁵. We hypothesised that these phenomena may be underpinned by
103 a progressively more precise and acoustically-invariant neural encoding of phonetic features across the first
104 year of life. This encoding would be expected to emerge as a neural response to speech that reflects a growing
105 invariance towards phonetic categories, where the limit case would be to have neural responses to phonetic
106 categories that are fully invariant to acoustic changes. This longitudinal investigation offers the first view into
107 phonetic feature encoding in the first year of life, while accounting for the full complexity of speech in
108 naturalistic listening environments. In the discussion, we connect our encoding analyses with previous work
109 on phonetic discrimination and consider the key role of perceptual invariance. Our results provide a
110 promising new avenue for developmental research with both infants and children. Precise measures of how
111 and when phonetic feature encoding evolves could serve as a complementary set of risk factors for
112 developmental language disorders, as well as illuminating the phonological trajectories experienced by both
113 typically- and atypically-developing children.

114

115 **Results**

116 **Robust neural tracking of acoustic and phonetic features in infants**

117 A multivariate TRF analysis was carried out to assess the low-frequency (1-15 Hz) neural encoding of speech
118 across the first year of life. Acoustic and phonetic features were extracted from the stimulus. Acoustic
119 features consisted of the 8-band acoustic spectrogram of speech (S) sound and the half-way rectified

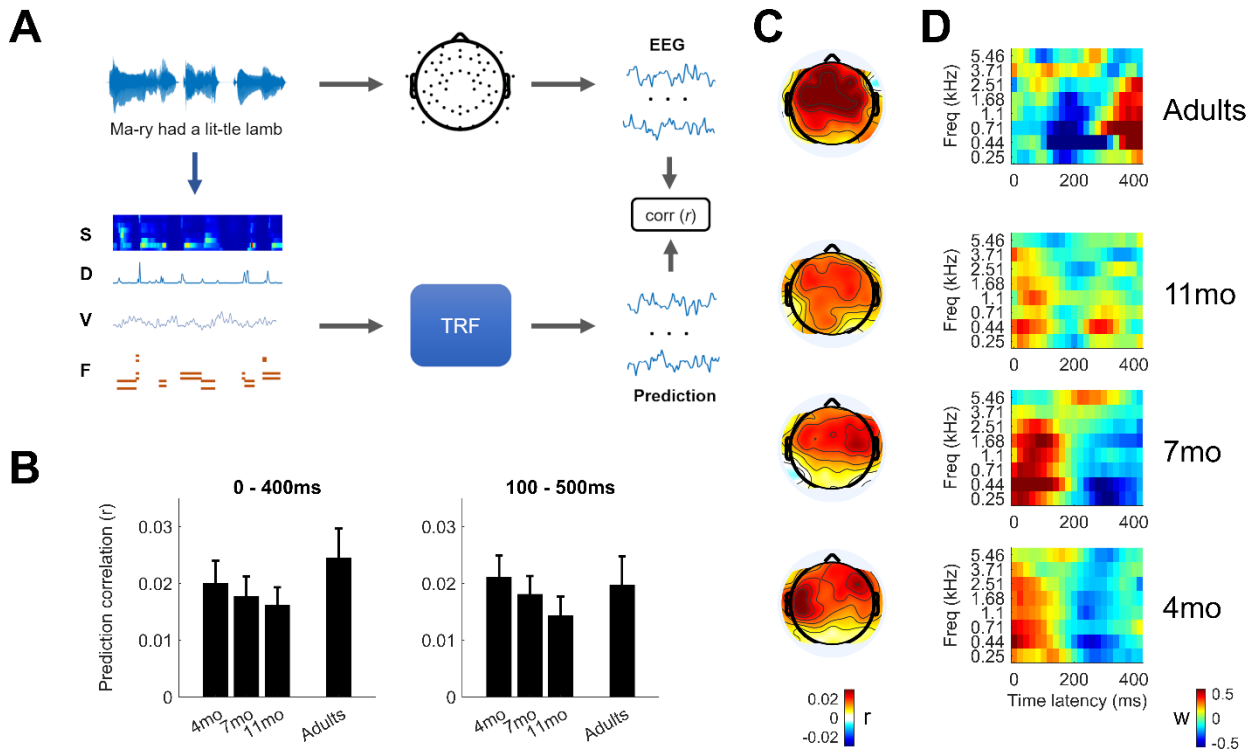
ⁱ This should not be intended as a hard boundary, as this is likely a gradual phenomenon that changes over large time windows, with differences between easy and more difficult speech contrasts

120 envelope derivative (D). Fourteen phonetic features were included to mark the categorical occurrence of
121 speech sounds, according to articulatory features describing voicing as well as manner and place of
122 articulation. To account for possible differences in the encoding of stressed and unstressed sounds, each
123 phonetic feature produced to two distinct vectors, leading to a 28-dimensional phonetic features matrix (F ;
124 see **Methods**). A nuisance regressor was also included to capture EEG variance related to visual motion (V).
125 Single-subject TRFs were derived for each experimental session to assess the cortical encoding of acoustic
126 and phonetic features by fitting a multivariate lagged regression model with all such features simultaneously
127 (**Figure 1A**).

128 EEG prediction correlations, calculated with leave-one-out cross-validation and averaged across all EEG
129 channels, were **greater than zero for all age groups** (one-sample Wilcoxon rank sum test, FDR-corrected for
130 multiple comparisons; 4mo: $p=9.5*10^{-6}$; 7mo: $p=9.5*10^{-6}$; 11mo: $p=9.5*10^{-6}$; adults: $p=2.2*10^{-4}$; **Figure 1B**).
131 Consistent with previous work, this analysis was carried out by considering speech-EEG lags from 0 to 400ms,
132 which were shown to largely capture the cortical acoustic-phonetic response in adults. The TRF analysis was
133 also repeated when considering a 100-500ms lag window, aiming to control for possible responses with
134 longer latencies in infants, while keeping the same model complexity (i.e., same window size). This analysis
135 also led to significant EEG predictions (one-sample Wilcoxon rank sum test, FDR-corrected; 4mo: $p=3.6*10^{-6}$;
136 7mo: $p=3.6*10^{-6}$; 11mo: $p=1.7*10^{-6}$; adults: $p=0.001$; **Figure 1B**), indicating a consistent speech-EEG
137 relationship involving acoustic and phonetic features in both time windows.

138 Topographic differences were expected both across participants and by age group due to major anatomical
139 changes during infancy³⁶. Larger EEG prediction correlations were measured in centro-frontal electrodes for
140 all age groups (**Figure 1C**), with **topographies becoming progressively more similar to those for adults** with
141 age in both the 0-400ms lag window (bootstrap with group size = 17 and 1000 iterations; average correlation
142 with adults: $r = 0.43, 0.51, 0.54$ for 4mo, 7mo, and 11mo respectively; repeated measures ANOVA on infant
143 data with age as the repeated factor: $F(2,1998) = 172.8, p = 6.1*10^{-70}$) and 100-500ms window (bootstrap;
144 average correlation with adults: $r = 0.32, 0.55, 0.52$ for 4mo, 7mo, and 11mo respectively; repeated measures

145 ANOVA: $F(2,1998) = 900.6$, $p = 1.5 \times 10^{-279}$). TRF models corresponding to spectrogram features are reported
 146 in **Figure 1D**, where weights were averaged across fourteen centro-frontal electrodes (25% of all channels)
 147 and all participants (see **Methods**).



148

149 **Figure 1: EEG tracking of acoustic and phonetic features in infants and adults.** (A) Schematic diagram of the
 150 analysis paradigm. Multivariate Temporal Response Function (TRF) models were fit to describe the forward
 151 relationship between speech features and the EEG signal recorded from adults and infants (4, 7, and 11mo).
 152 Speech features included the 8-band acoustic spectrogram (S), half-way rectified envelope derivative (D),
 153 visual motion (V), and phonetic features (F). (B) EEG prediction correlations of the multivariate TRF model
 154 were significant within each group for both the time-lag windows 0-400ms and 100-500ms. (C) Topographical
 155 patterns of the EEG prediction correlations in infants (shown for the TRF window 0-400ms) became
 156 progressively more similar to adults responses with age. (D) TRF weights corresponding to the S features
 157 averaged across centro-frontal electrodes.

158

159 Emergence of phonetic feature encoding in the first year of life

160 The analyses that follow aim to determine if and when cortical signals encode acoustically-invariant phonetic
 161 features during the first year of life. In line with previous behavioural work^{35,37-41} and current developmental
 162 theories^{1,34}, we expected categorical phonetic feature encoding to emerge from 6 months on (i.e., from the
 163 7mo recording session, in the present study), with progressively stronger encoding across the first year of life

164 visible by 11 months of age. To test this hypothesis (see H_{p2} in **Figure 2A**), phonetic feature encoding was
165 assessed based on the multivariate TRF models described in the previous section. Neural activity linearly
166 reflecting phonetic feature categories but not sound acoustics was accounted by subtracting EEG prediction
167 correlations corresponding to acoustic-only TRFs (which did not include phonetic features; see **Methods**)
168 from those corresponding to acoustic-phonetic TRFsⁱⁱ. For consistency with previous work^{27,29,32,33,42-44}, this
169 metric is referred to as FS-S (F: phonetic features; S: spectrogram and envelope derivative). As expected (H_{p1} -
170 α), FS-S values were progressively larger with age (average across all EEG channels; **Figure 2B**), with values
171 greater than zero emerging from 11 months of age for the time-latency window 0-400ms (one-sample
172 Wilcoxon rank sum test, FDR-corrected; 4mo: $p=0.237$; 7mo: $p=0.237$; **11mo: $p=0.037$; adults: $p=0.037$** ;
173 black bars indicate $p<0.05$), and **from 7 months** of age for the time-latency window 100-500ms (one-sample
174 Wilcoxon rank sum test, FDR-corrected; 4mo: $p=0.167$; **7mo: $p=0.044$; 11mo: $p=0.023$; adults: $p=0.044$**).
175 This latter finding is consistent with hypothesis 2 (H_{p2}), as depicted in **Figure 2A**. Furthermore, the 4mo group
176 did not show significant FS-S values for subsequent latency windows (200-600ms and 300-700ms).

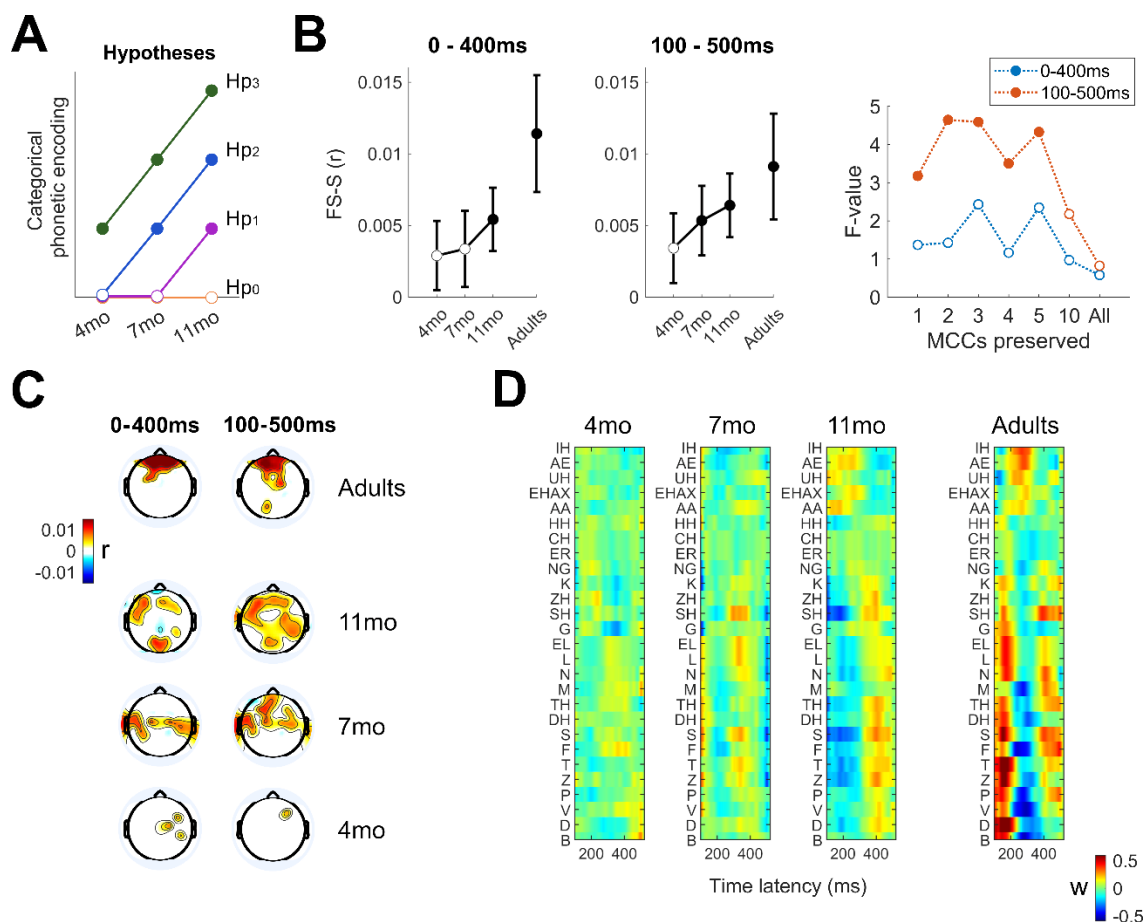
177 While the previous analysis identified significant phonetic encoding within individual age groups, the analysis
178 that follows explicitly assessed if phonetic encoding increased across the first year of life. TRF models and the
179 corresponding EEG prediction correlations showed large between-subject variability, which was expected
180 due to the noisy single-subject data. Under the assumption that participants in the same age group present
181 EEG responses to nursery rhymes with similar temporal patterns, a Multiway Canonical Correlation Analysis
182 (MCCA)⁴⁵ was carried out to isolate EEG components that are consistent within each group, substantially
183 improving the signal-to-noise ratio of the single-subject EEG. A repeated measures ANOVA test was then
184 carried out to determine if FS-S increased with age by considering the first MCCA component only (MCC_1)
185 i.e., the EEG component with highest temporal correlation across subjects within a given age group. A
186 significant increasing trend emerged for the 100-500ms window ($F(2,138)=3.19$, $p=0.044$) but not for the 0-

ⁱⁱ Please note that results did not change when acoustic-only TRFs consisted of acoustic vectors concatenated with shuffled phonetic information.

187 400ms latency window ($F(2,138)=1.37, p=0.257$; see coloured panel **Figure 2B**). The test was also run when
188 considering an increasing number of MCCs, showing significant results when considering up to five
189 components for the 100-500ms latency window. The EEG encoding of phonetic features was also studied at
190 individual electrodes, revealing robust encoding of phonetic features on large clusters of EEG channels in
191 adults as well as infants from 7 months of age, both when considering 0-400ms and 100-500ms windows
192 (**Figure 2C**; FDR-corrected one-sample Wilcoxon rank sum tests were run on each EEG channel; colours
193 indicate significant results with $p<0.05$).

194 Further analyses were carried out to assess the phonetic feature encoding at a fine-grained level, by studying
195 the TRF weights of the acoustic-phonetic TRF (weights are shown in **Figure 2D**). Phonetic distance maps³³
196 were calculated by using a multidimensional scaling analysis (MDS; see **Methods**) on the TRF weights
197 corresponding to phonetic features. This approach allows to quantify and visualise the level of similarity or
198 distance between datapoints by accounting simultaneously for multiple EEG channels and peri-stimulus time
199 latencies. By selecting the two most relevant MDS dimensions, the infant-adult Euclidean distance was
200 calculated for each age group (**Figure 3A**; bars indicate mean and SE calculated across 27 adult phonemes
201 computed as a linear combination of the corresponding phonetic features), showing that **the infant-adult**
202 **distance decreases with infant age** in the first year of life (repeated measures ANOVA, $F(2,52) = 18.2, p =$
203 $5.0 \cdot 10^{-7}$). *F*-score measures were derived quantifying the discriminability of specific phonetic feature groups
204 in the individual-subject TRFs using a *k*-means analysis (mean and SE were calculated on the *F*-scores resulting
205 from the 100 repetitions of *k*-means). The phonetic feature groupings considered for this analysis were place
206 of articulation, voicing, and manner of articulation (**Figure 3B,C**). As a validation step, the stability of the
207 resulting *F*-scores for the infants in the three longitudinal sessions was assessed over 100 repetitions of the
208 *k*-means procedure (repeated measures ANOVA, voicing: $F(2,198) = 155.8, p < 10^{-12}$; place of articulation:
209 $F(2,198) = 377.2, p < 10^{-12}$; manner of articulation: $F(2,198) = 29.4, p = 6.84 \cdot 10^{-12}$; **Figure 3B**). Next, statistical
210 analyses were carried out to determine the significance of the result across participants. As expected (due to
211 factors such as low-SNR, limited data, and inter-subject variability), single-subject phonetic feature maps did
212 not lead to significant results, even though results for place of articulation were trending towards significance

213 (repeated measures ANOVA, voicing: $F(2,92) = 0.6, p = 0.54$; place of articulation: $F(2,92) = 2.6, p = 0.08$;
 214 manner of articulation: $F(2,92) = 0.5, p = 0.61$). To compensate for the limited single-subject data, we ran a
 215 bootstrap analysis with 100 repetitions, each derived by averaging 17 subjects i.e., the same number of
 216 participants in the adult group. This analysis revealed that phonetic feature encoding increased with age for
 217 place of articulation and voicing, but not manner of articulation (repeated measures ANOVA, **voicing**:
 218 $F(2,198) = 21.3, p = 4.2 \cdot 10^{-9}$; **place of articulation**: $F(2,198) = 17.6, p = 9.0 \cdot 10^{-8}$; manner of articulation:
 219 $F(2,198) = 1.5, p = 0.22$).



220

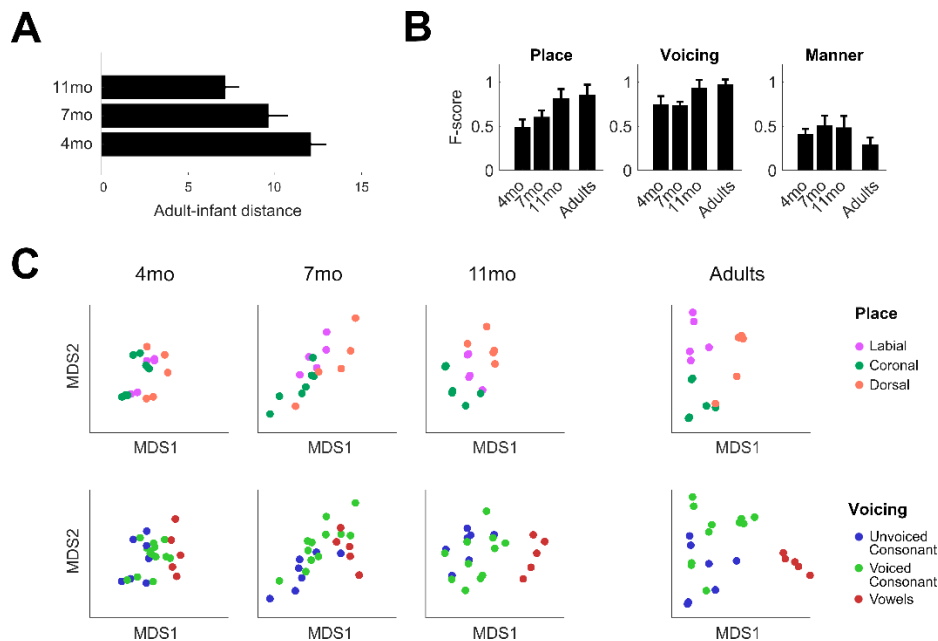
221 **Figure 2: Cortical encoding of phonetic features in the first year of life.** (A) *Hypotheses: The cortical encoding*
 222 *of phonetic feature categories was expected to emerge and progressively increase across the first year of life.*
 223 *Hypothesis 0 (Hp0): No phonetic encoding in the first year of life; Hp1-3: phonetic encoding from 11, 7, and 4*
 224 *months of age respectively.* (B) *Phonetic feature encoding measured as the EEG prediction correlation gain*
 225 *when including phonetic features in the TRF (mean and SE across participants, for the 0-400ms and 100-*
 226 *500ms lag windows). Black bars indicate significance ($p < 0.05$ after FDR-correction). The right panel indicates*
 227 *the F-statistics (repeated measures ANOVA) when using MCCA denoising (retaining 1, 2, 3, 4, 5, and 10*
 228 *components) and without MCCA denoising ('all'). A main effect of age emerged for the 100-500ms TRF when*
 229 *retaining up to 5 components (filled dots indicate significance; $p < 0.05$).* (C) *Phonetic feature encoding (EEG*

230 prediction correlation gain) across all electrodes. Coloured areas indicate significance ($p < 0.05$, t-test with FDR
231 correction). (D) TRF weights corresponding to phonetic features for the 100-500ms TRF.

232

233

234



235

236 **Figure 2: Sensitivity to phonetic feature groups in the first year of life.** (A) Distance between infant and adult
237 TRF weights (mean and SE). (B,C) Multidimensional scaling maps (MDS) were calculated on the phonetic
238 features TRFs as a function of peri-stimulus time lag and electrode. By carrying out 100 repeated k-means
239 classification, F-score measures were derived representing the discriminability of specific phonetic feature
240 groups in the TRFs (mean and SE across repetitions) i.e., place of articulation, voicing, and manner of
241 articulation (B). Individual MDS maps are shown in (C), where dots correspond to adult phonemes.

242

243 Discussion

244 The present investigation offers the first direct evidence that the human cortex encodes phonetic categories
245 during the first year of life, demonstrating significant phonetic encoding from 7 months of age and
246 progressively stronger encoding thereafter. A fine-grained and longitudinal understanding of the
247 development of phonetic feature encoding by the same infants listening to continuous speech was previously
248 absent from the literature. The behavioural and MMR infant speech processing literature has used targeted
249 experimental contrasts, focused largely on the perception of syllable stress and speech rhythm and on
250 phonetic category formation. As rhythm and stress patterns aid in identifying word boundaries, and phonetic

251 categories aid in comprehension (e.g., distinguishing ‘doggy’ from ‘daddy’), this prior work has been
252 important, showing that infants are sensitive to differences in speech rhythm from birth^{46,47}, and are sensitive
253 to some phonetic information as neonates⁴⁸. Nevertheless, no prior study has used continuous speech as a
254 basis for studying phonetic encoding. Consequently, our findings have several implications for understanding
255 of the development of speech processing.

256 Currently, it remains unclear how and at what stage of development phonetic category encoding is learnt.
257 This question remains open largely because of methodological constraints. The present study offers, for the
258 first time, direct evidence on the ‘when’ of phonetic category learning. There is a consensus in the literature
259 that discriminating phonetic categories is a key processing step regarding speech comprehension by adults⁴⁹,
260 although see Feldman et al., 2021⁵⁰ for recent caveats regarding infants. While adult studies used direct
261 invasive recordings to measure the cortical encoding of phonetic categories⁵¹, recent methodological
262 developments (i.e., the TRF framework^{7,8}) allowed us to circumvent some of the major challenges
263 encountered by previous infant studies, thereby providing more precise developmental information.

264 The assessment of phonetic encoding as operationalised here fulfils three main elements of novelty that go
265 beyond any previous investigation. First, we studied the cortical encoding of phonetic categories in infants
266 with **direct neural measurements** based on EEG and as part of an unprecedented targeted longitudinal
267 investigation. Second, the use of the forward TRF framework allowed us to assess phonetic category
268 **encoding**, rather than relying on the typical sound discrimination metrics used in prior behavioural^{35,37-41} and
269 neurophysiology studies (e.g., MMR)^{4,52-56}. Third, the TRF framework allowed us to study the perception of
270 **natural speech** in infants, instead of focusing on selected phonetic or word contrasts, as in the past literature.
271 This is a crucial step forward, as the discriminatory skills that infants exhibit in simplified laboratory settings
272 (e.g., isolated syllable discrimination measured via a head-turn or looking procedure or with MMRs) may not
273 be sufficient for detecting phonetic categories in naturalistic settings.

274 The present study indicates that phonetic category encoding during natural speech listening emerges
275 between 5 and 7 months of age. This provides the literature with new and fundamental insights into the

276 development of speech processing in neurotypical infants. Further, the TRF approach yields novel
277 information on *which* specific phonetic contrasts evolve with age, demonstrating a natural progression
278 toward adult phonetic encoding. This insight is further reinforced by the observation that acoustic encoding
279 did not increase with age. Consequently, the enhanced phonetic encoding with age observed here could not
280 simply be due to stronger acoustic encoding, as acoustic encoding showed the opposite pattern (a non-
281 significant decreasing trend with age). Interestingly, our results suggest that 4mo pre-babbling infants,
282 despite being equipped with the fundamental combinatorial code for speech analysis⁵⁷, do not yet exhibit
283 categorical phonetic encoding. Based on these results, we can speculate that prior demonstrations of infant
284 behavioural and MMR discrimination between syllables like "pa" and "ba" probably have an acoustic basis
285 but do not reflect categorical phonetic encoding. In other words, the ability to distinguish two sounds does
286 not necessarily mean that those sounds are encoded as separate categories. Our study is instead probing
287 that categorical encoding directly.

288 One challenge with longitudinal neurophysiology studies in infants is the substantial anatomical change that
289 occurs with age, meaning that while macroscopic patterns are likely to remain consistent (e.g., temporal vs.
290 occipital), there cannot be a channel-by-channel correspondence between age groups, even when
291 considering the same participants. For this reason, the majority of this investigation focused on measures
292 combining multiple EEG channels simultaneously (e.g., **Figure 1D** was an average of 14 centro-frontal
293 channels). These considerations make the topographical distribution of phonetic encoding strength shown
294 in **Figure 2C** even more remarkable, as a centro-frontal cluster of EEG channels was shown to reflect phonetic
295 encoding across all age groups with the exception of 4 months, where phonetic encoding as assessed by the
296 TRF was not significant.

297 Our phonetic encoding results showed topographical patterns and TRF weights for adults that differ from the
298 prior adult EEG literature on natural speech listening TRFs^{27,29}. While part of the discrepancy may be due to
299 the use of a different EEG acquisition device and to the use of audio-visual stimuli, the primary explanation
300 is likely to be the choice of stimuli. This is the first TRF investigation of phonetic processing with adults

301 involving a nursery rhyme listening task. Nursery rhymes are indeed a form of natural speech which is more
302 suited to infants. The rhythmic cues and exaggerated stress patterns characterising nursery rhymes have
303 been demonstrated to be important elements supporting speech perception and language learning^{58,59},
304 accordingly they were ideal stimuli for the Cambridge UK BabyRhythm study. In prior TRF work, we have
305 demonstrated similar envelope entrainment to these nursery rhymes by adults and infants²⁶. Nevertheless,
306 it is important to note that the regular rhythms and melodic properties of nursery rhymes makes the different
307 from the typical speech TRF stimuli used with adults, such as audio-books and podcasts. As such, the TRF
308 results were expected to show different spatio-temporal patterns for adult listeners compared to previous
309 TRF work.

310 The results of this study add to the growing literature on cortical speech tracking^{21,27,31,33,44,60-62}. While the
311 literature typically focuses on the cortical tracking of the speech envelope^{16,24,63-66} (including previous
312 analyses of this dataset^{21,26}), the present investigation enriches our understanding of phonetic feature TRFs.
313 Prior TRF studies of phonetic encoding in adults and children have revealed that phonetic processing is
314 affected by speech clarity⁴³, selective attention²⁹, and proficiency in a second language³³, and shows
315 correlations with psychometric measures of phonemic awareness³². The present study demonstrates that
316 emergent phonetic TRFs can also be measured in pre-verbal infants, providing a novel window into infant
317 perception and cognition. Whilst recent developments have started to use neural tracking to predict
318 language development in infants⁶⁷, further research will also determine whether a robust relationship exists
319 between speech TRFs and other related aspects of cognition (e.g., selective attention, prediction) in infants,
320 and when such related aspects come on-line. Further research with infants at family risk for disorders of
321 language learning may also reveal when and how developmental trajectories are impacted by developmental
322 disorders that are carried genetically, such as developmental dyslexia and developmental language disorder.
323 Such work could be very valuable regarding early detection and improved mechanistic understanding of
324 these disorders.

325 In summary, this study demonstrated the emergence of phonetic encoding from 7 months of age using direct
326 neural measurements during natural speech listening. The data provide clear-cut evidence of the emergence
327 of phonetic categories that contributes to the current debate regarding their role in the development of
328 speech processing. Our demonstration that phonetic encoding can be assessed with nursery rhyme stimuli
329 in ecologically-valid conditions opens the door to cross-language work using TRFs that investigates the
330 interaction between characteristics of natural language such as phonological complexity and the
331 development of phonetic encoding. It also provides opportunities for novel mechanistic investigations of the
332 development of bi-lingual and multi-lingual lexicons during language acquisition.

333

334

335 Online Methods

336 Subjects and experimental procedure

337 The present study carried out a re-analysis of an EEG dataset involving a speech listening task in a longitudinal
338 cohort of fifty infants (first part of a larger cohort of 122 subjects²¹). Participants were infants born full term
339 (37-42 gestational weeks) and had no diagnosed developmental disorder, recruited from a medium sized city
340 in the United Kingdom and surrounding areas via multiple means (e.g., flyers in hospitals, schools, and
341 antenatal classes, research presentations at maternity classes, online advertising). The study was approved
342 by the Psychology Research Ethics Committee of the University of Cambridge. Parents gave written informed
343 consent after a detailed explanation of the study and families were repeatedly reminded that they could
344 withdraw from the study at any point during the repeated appointment. The experiment involved three EEG
345 recording sessions when the infants (24 male and 26 female) were 4 months old (4mo; 115.6 ± 5.3 days), 7
346 months old (7mo; 212.5 ± 7.2 days) and 11 months old (11mo; 333.0 ± 5.5 days) [mean \pm standard deviation
347 (*SD*)]. A bilingualism questionnaire (collected from 45 out of the 50 infants) ascertained that 38 of the infants
348 were exposed to a monolingual environment and 12 were exposed multilingual environment, of these 93.5%
349 (43 infants) reported English as the primary language exposed to the infant. Note that this was a longitudinal
350 investigation, meaning that the same 50 infants were tested at 4, 7, and 11 months of age. In addition to the
351 150 EEG sessions from the infant dataset, this study also analysed EEG data from twenty-two monolingual,
352 English-speaking adult participants performing the same listening task (11 male, aged 18-30, mean age: 21).
353 Data from four adult participant was excluded due to inconsistencies with the synchronisation triggers,
354 leaving seventeen participants data for the analysis.

355 Infant participants were seated in a highchair (one metre in front of their primary caregiver) in a sound-proof
356 acoustic chamber, while adult participants were seated in a normal chair. All participants were seated 650mm
357 away from the presentation screen. EEG data were recorded at a sampling rate of 1 kHz using a GES 300
358 amplifier using a Geodesic Sensor Net (Electrical Geodesics Inc., Eugene, OR, United States). 64 and 128
359 channels were used for infants and adults respectively. Sounds were presented at 60 dB from speakers placed
360 either side of the screen (Q acoustics 2020i driven by a Cambridge Audio Topaz AM5 Stereo amplifier).
361 Participants were presented with eighteen nursery rhyme videos played sequentially, each repeated 3 times
362 (54 videos with a presentation time of 20' 33" in total). Adult participants were asked to attend to the audio-
363 visual stimulus while minimising their motor movements. All adult participants completed the full
364 experiment. Infants listened to at least two repetitions of each nursery rhyme (minimum of 36 nursery
365 rhymes lasting 13' 42"). The experiment included other elements that were not relevant to the present study
366 (e.g., resting state EEG; please refer to the previous papers on this dataset for further information^{21,26}).

367

368 Stimuli

369 A selection of eighteen typical English language nursery rhymes was chosen as the stimuli. Audio-visual
370 stimuli of a singing person (upper-body only) were recorded using a Canon XA20 video camera at 1080p,
371 50fps and with audio at 4800 Hz. A native female speaker of British English used infant-directed speech to
372 melodically sing (for example "Mary Quite Contrary") or rhythmically chant (for nursery rhymes like "There
373 was an old woman who lived in a shoe") the nursery rhymes whilst listening to a 120 bpm metronome

374 through an intra-auricular headphone (e.g., allowing for 1Hz and 2Hz beat rates; see Figs. S2 and S4 from
375 Attaheri et al.²¹). The metronome's beat was not present on the stimulus audios and videos, but it ensured
376 that a consistent rhythmic production was maintained throughout the 18 nursery rhymes. To ensure natural
377 vocalisations, the nursery rhyme videos were recorded sung, or rhythmically chanted, live to an alert infant.
378

379 **Data preprocessing**

380 Analyses were conducted with MATLAB 2021a by using custom scripts developed starting from publicly
381 available scripts shared by the CNSP initiative (Cognition and Natural Sensory Processing;
382 <https://cnspsworkshop.net>; see section Data and Code Availability for further details).

383 In order to carry out the same preprocessing and analysis pipeline on infants and adult EEG data, the adult
384 128-channel EEG data was transformed into a 64-channel dataset via spline interpolation, with the relative
385 channel locations corresponding to those of the infant participants. All subsequent analyses on infants and
386 adult were identical.

387 The four facial electrodes (channels 61-64) were excluded from all analyses, as they are not part of the
388 specific infant-sized EGI Geodesic sensor net. The EEG data from the remaining 60 channels was band-pass
389 filtered between 1 and 15 Hz by means of zero-phase shift Butterworth filters with order 2 (by using the
390 filtering functions in the CNSP resources). EEG signals were downsampled to 50 Hz. Next, Artifact Subspace
391 Reconstruction (ASR; *clean_asr* function from EEGLAB⁶⁸) was used to clean noise artifacts from the EEG
392 signals. Channels with excessive noise (which could not be corrected with ASR) were identified via probability
393 and kurtosis and were interpolated via spherical interpolation, if they were three standard deviations away
394 from the mean. EEG signals were then re-referenced to the average of the two mastoid channels, which were
395 then removed from the data, producing a preprocessed EEG dataset with 58 channels. Data from repeated
396 trials was then averaged. Three infant subjects were removed because of excessive noise in at least one of
397 their three recording sessions.

398

399 **Sung speech representations**

400 The present study involved the measurement of the coupling between EEG data and various properties of
401 the sung speech stimuli. These properties were extracted from the stimulus data based on methodologies
402 developed in previous research. First, we defined a set of descriptors summarising low-level acoustic
403 properties of the speech stimuli. Acoustic features consisted of an 8-band acoustic spectrogram (S) and a
404 half-way rectified broadband envelope derivative (D)^{33,60}. S was obtained by filtering the sound waveform
405 into eight frequency bands between 250 and 8 kHz that were logarithmically spaced according to the
406 Greenwood equation⁶⁹. The broadband envelope was calculated as the sum across the eight frequency bands
407 of S. The D signal was then derived by calculating the derivative of the broadband envelope, and by half-way
408 rectifying the resulting signal. Second, fourteen phonetic features were then selected to mark the categorical
409 occurrence of speech sounds, according to articulatory features describing voicing, manner, and place of
410 articulation^{70,71}: voiced consonant, unvoiced consonant, plosive, fricative, nasal, strident, labial, coronal,
411 dorsal, anterior, front, back, high, low. To account for possible differences in the encoding of stressed and
412 unstressed sounds, each phonetic feature was assigned to two distinct vectors, leading to a 28-dimensional
413 phonetic features matrix (F). The precise timing of the phonetic units was identified in three steps. First,

414 syllable and phoneme sequences were obtained from the transcripts of the nursery rhymes. Second, an initial
415 alignment was derived by identifying the syllabic rate and syllable onsets for each piece, and then assigning
416 the phonemes in a syllable starting from the corresponding onset time. This automatic alignment was stored
417 according to the TextGrid format⁷². Third, the phoneme alignments were manually adjusted using Praat
418 software⁷². Phonetic feature vectors were produced in MATLAB software to categorically mark the
419 occurrence of phonetic units from start to finish with unit rectangular pulses²⁷. Finally, a nuisance regressor
420 was also included to capture EEG variance related with visual motion (V), which was derived as the frame-to-
421 frame luminance change, averaged across all pixels.

422

423 **Multivariate Temporal Response Function (mTRF)**

424 A single input event at time t_0 affects the neural signals for a certain time window $[t_1, t_1+t_{win}]$, with $t_1 \geq 0$ and
425 $t_{win} > 0$. Temporal response functions (TRFs) describe this relationship at the level of individual subject and
426 EEG channel. In this study, TRFs were estimated by means of a multivariate lagged regression, which
427 determines the optimal linear transformation from stimulus features to EEG (forward model)^{13,73}. A
428 multivariate TRF model (mTRF) was fit for each subject by considering all features simultaneously (S, D, F,
429 and V; **Figure 1A**) with the mTRF-Toolbox^{7,8}. While a time-lag window of 0-400 ms was considered sufficient
430 to largely capture the acoustic-phonetic/EEG relationship with a single-speaker listening task in adults, based
431 on previous studies (e.g., Di Liberto et al.²⁷), the relevant latencies in infants were unknown. To account for
432 possible slower or delayed response in infants, a time-latency window of 100-500 ms (i.e., same duration but
433 with longer latency) was also included in the analysis for all groups. The reliability of the TRF models was
434 assessed using a leave-one-out cross-validation procedure (across trials i.e., nursery rhymes), which
435 quantified the EEG prediction correlation (Pearson's r) on unseen data while controlling for overfitting. The
436 TRF model calculation included a Tikhonov regularization, which involves the tuning of a regularization
437 parameter (λ) that was conducted by means of an exhaustive search of a logarithmic parameter space from
438 0.01 to 10^6 on the training fold of each cross-validation iteration^{7,8}. Note that the correlation values were
439 calculated with the noisy EEG signal; therefore, the r -scores could be highly significant even though they have
440 low absolute values ($r \sim 0.1$ for sensor-space low-frequency EEG^{27,30,60}).

441

442 **Phonetic distance maps**

443 We sought to study the effect of development on phonetic perception by projecting the TRF weights
444 corresponding with the phonetic features onto a space in which distance represents the perceptual
445 separation between phonological units. The TRF weights for the phonetic features, which were represented
446 in a 28-dimensional space, were projected to the phoneme space. To do so, the TRF for a given phoneme
447 was calculated as the sum of the TRF weights of the corresponding features. This produced a 54-dimensional
448 matrix: 27 stressed and 27 unstressed phonemes. The weights corresponding to the two versions of a
449 phoneme (stressed and unstressed) were then combined when projecting to the phonetic feature map,
450 obtaining a 27-dimensional space. Specifically, a classical multidimensional scaling (MDS) was used to project
451 the phoneme TRF weights (phonemes were considered as objects and time latencies were considered as
452 dimensions) onto a multidimensional space for each age group, in which distances represented the
453 discriminability of particular phonetic contrasts in the EEG signal. The result for each infant group was then
454 mapped to the average adult MDS space by means of a Procrustes analysis (MATLAB function *procrustes*).
455 This analysis allowed us to project the infant phonetic feature maps for different proficiency levels to a

456 common multidimensional space where they could be compared quantitatively; we call these maps phonetic
457 feature distance maps. Note that while this transformation does not assume nor imply categorical phonemic
458 encoding, it indeed allows to quantify and visualise the encoding of phonetic features across age, by using
459 familiar phonemic units.

460 While **Figure 3A** shows the infant-adult distances in the MDS space, the effect of age on phonetic feature
461 encoding was also quantified with a clustering analysis. Specifically, the randomised clustering algorithm *k*-
462 means was run on the MDS maps for each group to determine whether the phonetic feature groups could
463 be deduced from the data without supervision. We performed 100 repetitions of *k*-means and a classification
464 *F*-Score (or *F*₁-Score) was calculated on each of them, obtained as the harmonic mean of precision and recall.
465 We performed this procedure for the three feature sets of interest: a) the three-class feature-set with
466 'vowels', 'voiced consonants', and 'unvoiced consonants'; b) the three-class feature-set describing the place
467 of articulation, with features 'labial', 'coronal', and 'dorsal'; and c) the three-class feature-set describing the
468 manner of articulation, with features 'fricative', 'stop', and 'nasal'. Since all feature-sets were three-
469 dimensional, *k* had always value three. Note that *k*-means is an unsupervised clustering algorithm, meaning
470 that there was no direct correspondence between the clusters and the classes of interest. As such, *F*-scores
471 were selected for the best matching assignment of classes on each execution of *k*-means. A large *F*-score
472 corresponded to a strong encoding of the feature-set of interest in the EEG data.

473

474 **Multiway Canonical Correlation Analysis (MCCA)**

475 EEG data is notorious for its low signal-to-noise ratio (SNR), which represent one of the core challenges when
476 analysing this kind of data. One approach to improve the SNR is multiway canonical correlation analysis
477 (MCCA), a tool that identifies EEG components that are most correlated across subjects. Under the
478 assumption that a stimulus would produce consistent cortical responses across subjects in the same age-
479 group, MCCA identifies such consistent responses accepting that they may originate from distinct sources
480 (i.e., distinct topographical patterns) for different subjects. MCCA is an extension of canonical correlation
481 analysis⁷⁴ to the case of multiple (> 2) subjects. Given *N* multichannel datasets *Y*_{*i*} with size *T* × *J*_{*i*}, 1 ≤ *i* ≤ *N* (time
482 × channels), MCCA finds a linear transform *W*_{*i*} (sizes *J*_{*i*} × *J*₀, where *J*₀ < min(*J*_{*i*})_{1 ≤ *i* ≤ *N*}), which, when applied to the
483 corresponding data matrices, aligns them to common coordinates and reveals shared patterns⁴⁵. These
484 patterns can be derived by summing the transformed data matrices as follows: $\sum_{i=1}^N Y_i W_i$. The columns of
485 the matrix *Y*, which are mutually orthogonal, are referred to as summary components (SC). The first
486 components are signals that most strongly reflect the shared information across the several input datasets,
487 thus minimising subject-specific and channel-specific noise. Here, MCCA was run within each age group (4mo,
488 7mo, 11mo, and adults). After fitting the MCCA mapping and projecting the data to the SC space, a given
489 number of component was retained (e.g., only the first component) before performing the inverse mapping
490 and obtain a denoised version of the EEG signal for each subject. This denoising procedure was repeated by
491 retaining a progressive number of components (**Figure 2B**).

492

493 **Statistical analysis**

494 All statistical analyses directly comparing the groups were performed using a repeated measures ANOVA.
495 One-sample Wilcoxon signed-rank tests were used for post hoc tests. Correction for multiple comparisons
496 was applied where necessary via the false discovery rate (FDR) approach. In that case, the FDR adjusted *p*-

497 value was reported. Descriptive statistics for the neurophysiology results are reported as a combination of
498 mean and standard error (SE).
499

500 **Data and code availability**

501 Analyses were conducted by using custom MATLAB scripts developed starting from publicly available scripts
502 shared by the CNSP initiative (Cognition and Natural Sensory Processing; <https://cnspsworkshop.net>). Such
503 analysis scripts avail of external publicly available libraries: the mTRF-Toolbox
504 (<https://github.com/mickcrosse/mTRF-Toolbox>)⁸, EEGLAB⁶⁸; and the NoiseTools library
505 (<http://audition.ens.fr/adc/NoiseTools>)⁴⁵. Data was converted to the CNL data structure (Continuous-event
506 Neural Data - <https://cnspsworkshop.net>), allowing to carry out the analyses with the CNSP analysis scripts,
507 which provided a platform for bringing together all the necessary libraries. We commit to publicly share the
508 EEG data in the first half of 2023. Study data were collected and managed using REDCap (Research Electronic
509 Data Capture) electronic data capture tools hosted at Cambridge university^{75,76}.

510 **References**

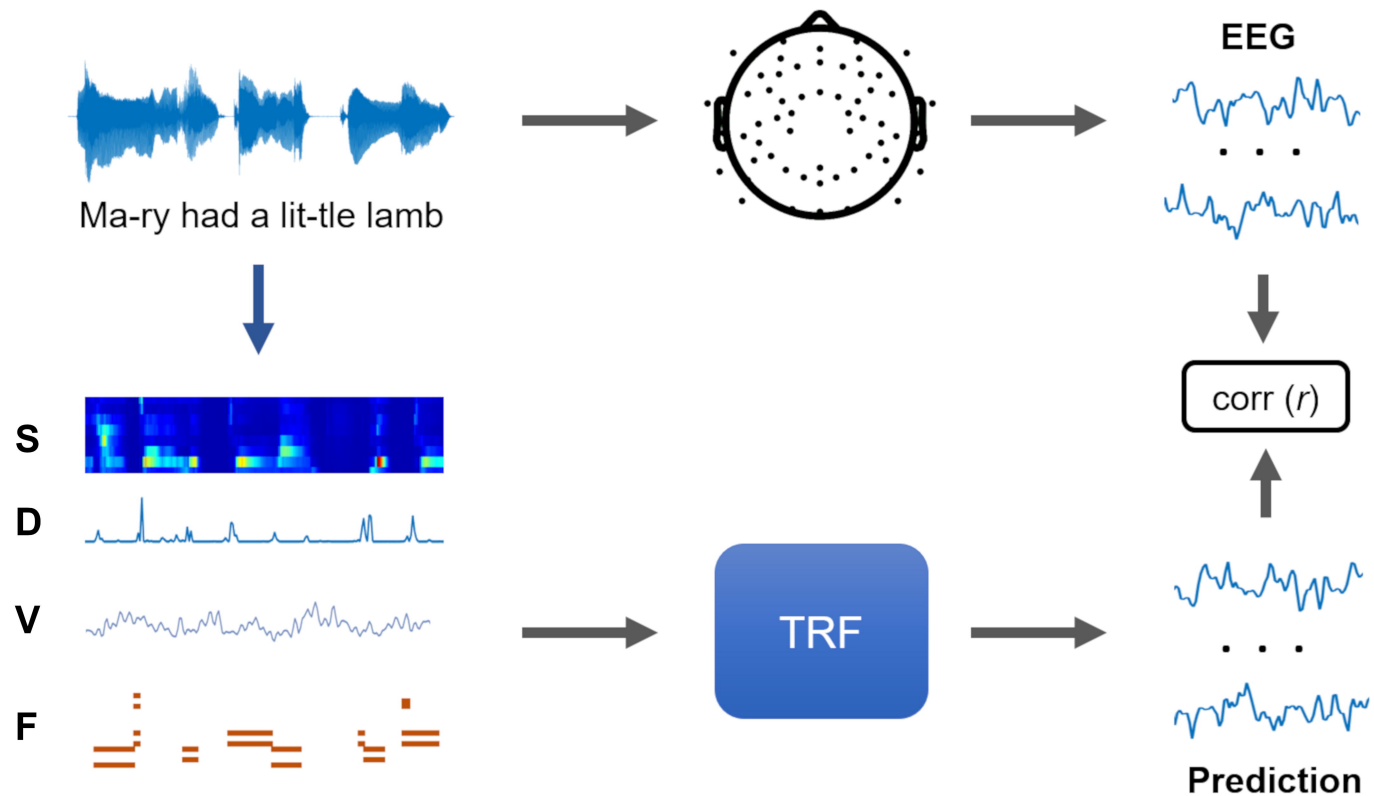
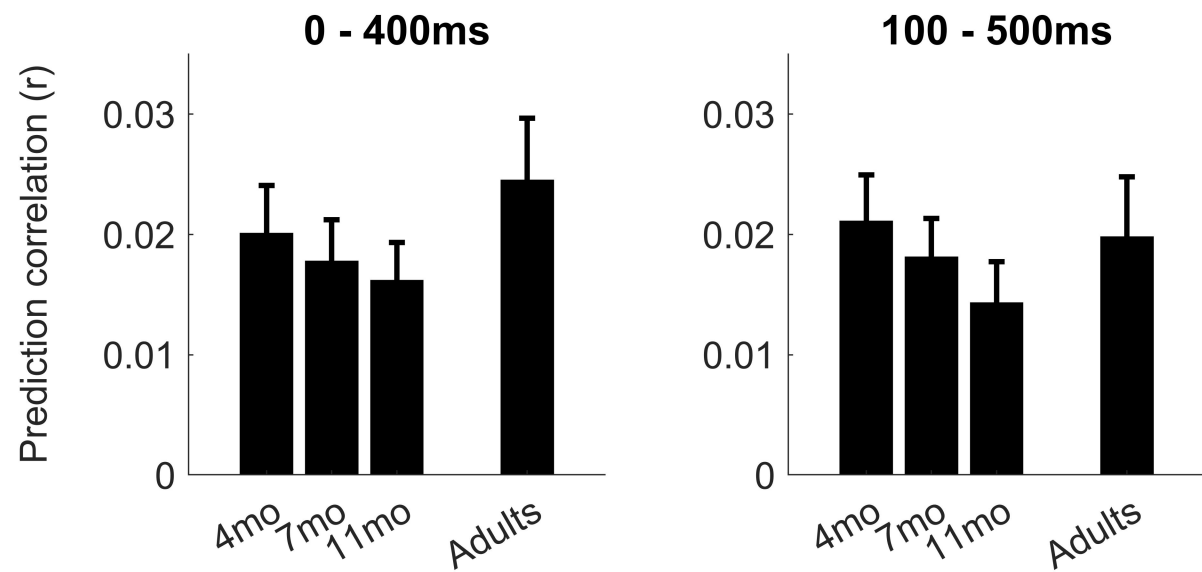
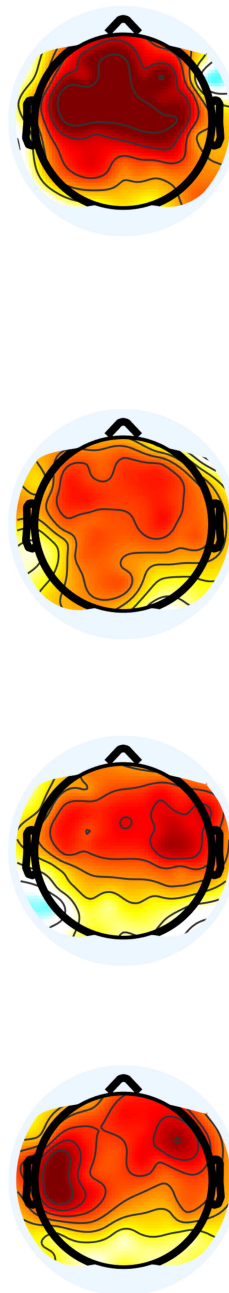
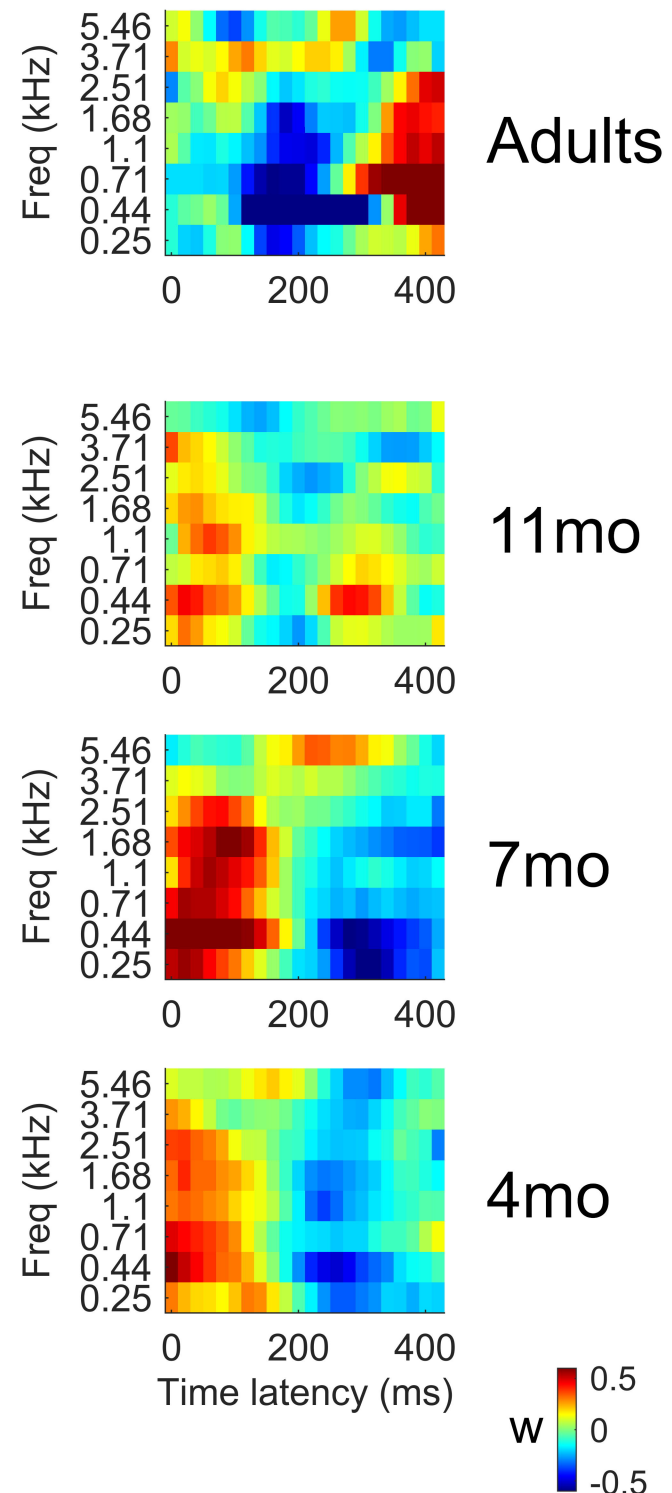
- 511 1. Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. *Nat Rev Neurosci* 5, 831-843.
512 10.1038/nrn1533.
- 513 2. Kuhl, P., and Rivera-Gaxiola, M. (2008). Neural Substrates of Language Acquisition. *Annual Review of*
514 *Neuroscience* 31, 511-534. 10.1146/annurev.neuro.30.051606.094321.
- 515 3. Kuhl, P.K. (2010). Brain mechanisms in early language acquisition. *Neuron* 67, 713-727.
516 10.1016/j.neuron.2010.08.038.
- 517 4. Wu, Y.J., Hou, X., Peng, C., Yu, W., Oppenheim, G.M., Thierry, G., and Zhang, D. (2022). Rapid learning
518 of a phonemic discrimination in the first hours of life. *Nature Human Behaviour* 6, 1169-1179.
519 10.1038/s41562-022-01355-1.
- 520 5. Dehaene-Lambertz, G., and Gliga, T. (2004). Common neural basis for phoneme processing in infants
521 and adults. *J Cogn Neurosci* 16, 1375-1387. 10.1162/0898929042304714.
- 522 6. Csibra, G., Kushnerenko, E., and Grossmann, T. (2008). Electrophysiological methods in studying
523 infant cognitive development.
- 524 7. Crosse, M.J., Zuk, N.J., Di Liberto, G.M., Nidiffer, A.R., Molholm, S., and Lalor, E.C. (2021). Linear
525 Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological
526 Considerations for Applied Research. *Frontiers in neuroscience* 15, 705621-705621.
527 10.3389/fnins.2021.705621.
- 528 8. Crosse, M.J., Di Liberto, G.M., Bednar, A., and Lalor, E.C. (2016). The multivariate temporal response
529 function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli.
530 *Frontiers in Human Neuroscience* 10. 10.3389/fnhum.2016.00604.
- 531 9. Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in*
532 *Psychology* 25, 143-170. <https://doi.org/10.1016/j.newideapsych.2007.02.001>.
- 533 10. Ziegler, J.C., and Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled
534 reading across languages: a psycholinguistic grain size theory. *Psychol Bull* 131, 3-29. 10.1037/0033-
535 2909.131.1.3.
- 536 11. Obleser, J., and Kayser, C. (2019). Neural Entrainment and Attentional Selection in the Listening Brain.
537 *Trends in Cognitive Sciences*. Elsevier Ltd.
- 538 12. Lalor, E.C., and Foxe, J.J. (2010). Neural responses to uninterrupted natural speech can be extracted
539 with precise temporal resolution. *European Journal of Neuroscience* 31, 189-193. 10.1111/j.1460-
540 9568.2009.07055.x.

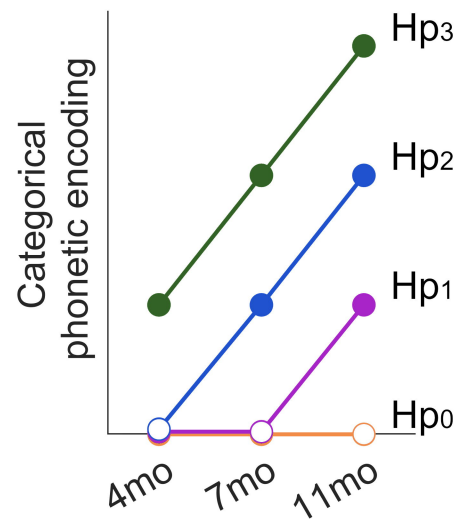
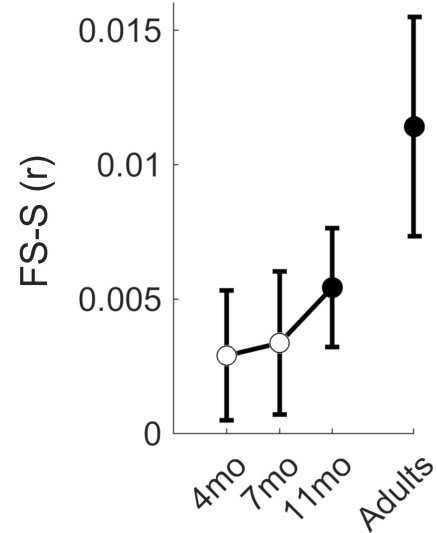
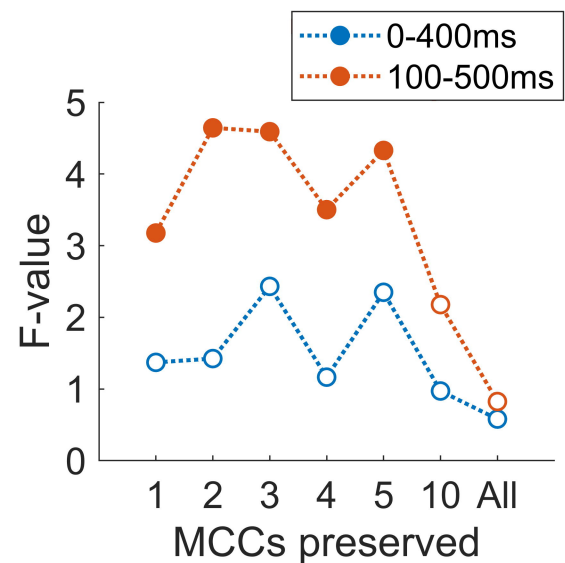
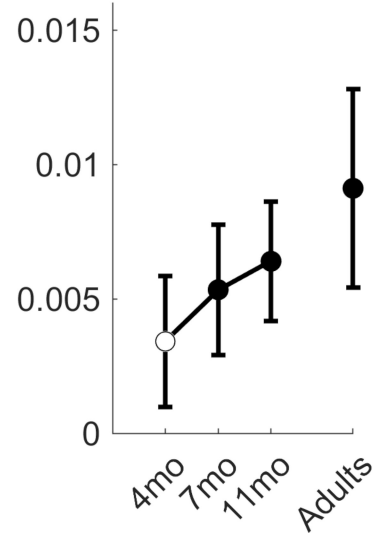
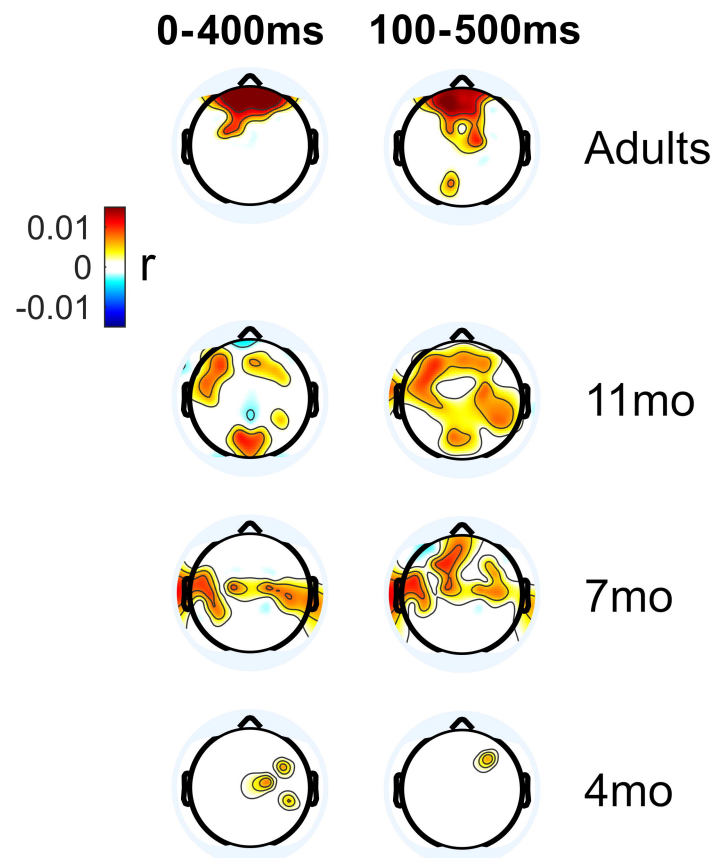
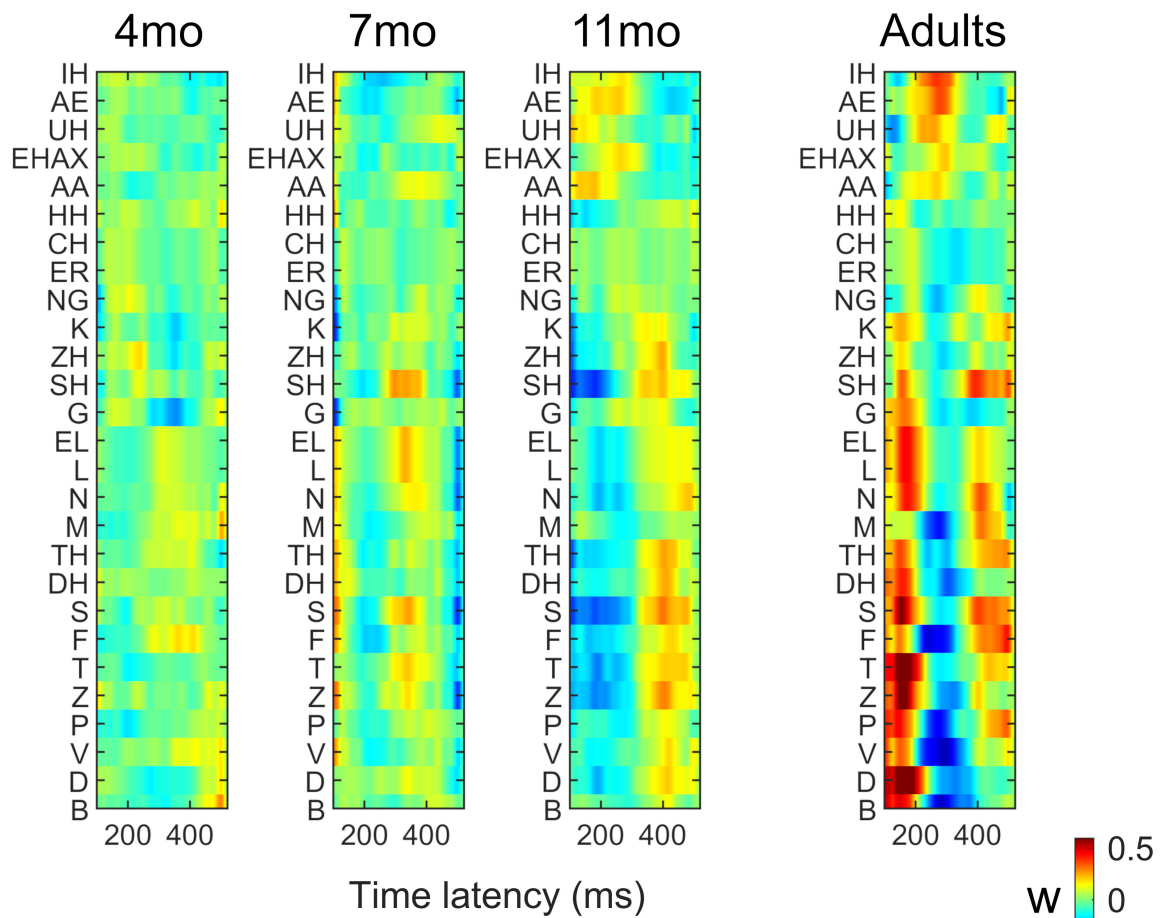
- 541 13. Lalor, E.C., Power, A.J., Reilly, R.B., and Foxe, J.J. (2009). Resolving Precise Temporal Processing
542 Properties of the Auditory System Using Continuous Stimuli. *Journal of Neurophysiology* *102*, 349-
543 359. 10.1152/jn.90896.2008.
- 544 14. Aiken, S.J., and Picton, T.W. (2008). Human cortical responses to the speech envelope. *Ear and*
545 *Hearing* *29*, 139-157.
- 546 15. Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to
547 competing speakers. *Proc Natl Acad Sci U S A* *109*, 11854-11859. 10.1073/pnas.1205381109.
- 548 16. O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney,
549 M., Shamma, S.A., and Lalor, E.C. (2014). Attentional Selection in a Cocktail Party Environment Can
550 Be Decoded from Single-Trial EEG. *Cerebral Cortex*, bht355-bht355.
- 551 17. Mesgarani, N., and Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-
552 talker speech perception. *Nature* *485*, 233-U118. 10.1038/nature11020.
- 553 18. Hjortkjaer, J., Märcher-Rørsted, J., Fuglsang, S.A., and Dau, T. (2020). Cortical oscillations and
554 entrainment in speech processing during working memory load. *Eur J Neurosci* *51*, 1279-1289.
555 10.1111/ejn.13855.
- 556 19. Leonard, M.K., Baud, M.O., Sjerps, M.J., and Chang, E.F. (2016). Perceptual restoration of masked
557 speech in human cortex. *Nature Communications* *7*, 13619-13619. 10.1038/ncomms13619
558 <http://www.nature.com/articles/ncomms13619#supplementary-information>.
- 559 20. Di Liberto, G.M., Lalor, E.C., and Millman, R.E. (2018). Causal cortical dynamics of a predictive
560 enhancement of speech intelligibility. *NeuroImage* *166*. 10.1016/j.neuroimage.2017.10.066.
- 561 21. Attaheri, A., Choidealbha Á, N., Di Liberto, G.M., Rocha, S., Brusini, P., Mead, N., Olawole-Scott, H.,
562 Boutris, P., Gibbon, S., Williams, I., et al. (2022). Delta- and theta-band cortical tracking and phase-
563 amplitude coupling to sung speech by infants. *Neuroimage* *247*, 118698.
564 10.1016/j.neuroimage.2021.118698.
- 565 22. Jessen, S., Fiedler, L., Münte, T.F., and Obleser, J. (2019). Quantifying the individual auditory and
566 visual brain response in 7-month-old infants watching a brief cartoon movie. *NeuroImage* *202*,
567 116060-116060. 10.1016/j.neuroimage.2019.116060.
- 568 23. Jessica Tan, S.H., Kalashnikova, M., Di Liberto, G.M., Crosse, M.J., and Burnham, D. (2022). Seeing a
569 talking face matters: The relationship between cortical tracking of continuous auditory - visual
570 speech and gaze behaviour in infants, children and adults. *NeuroImage* *256*, 119217.
571 <https://doi.org/10.1016/j.neuroimage.2022.119217>.
- 572 24. Kalashnikova, M., Peter, V., Di Liberto, G.M., Lalor, E.C., and Burnham, D. (2018). Infant-directed
573 speech facilitates seven-month-old infants' cortical tracking of speech. *Scientific Reports* *8*.
574 10.1038/s41598-018-32150-6.
- 575 25. Ortiz Barajas, M.C., Guevara, R., and Gervain, J. (2021). The origins and development of speech
576 envelope tracking during the first months of life. *Developmental Cognitive Neuroscience* *48*, 100915.
577 <https://doi.org/10.1016/j.dcn.2021.100915>.
- 578 26. Attaheri, A., Panayiotou, D., Phillips, A., Choidealbha, Á.N., Di Liberto, G.M., Rocha, S., Brusini, P.,
579 Mead, N., Flanagan, S., and Olawole-Scott, H. (2022). Cortical Tracking of Sung Speech in Adults vs
580 Infants: A Developmental Analysis. *Frontiers in neuroscience* *16*.
- 581 27. Di Liberto, G.M., O'Sullivan, J.A., and Lalor, E.C. (2015). Low-frequency cortical entrainment to speech
582 reflects phoneme-level processing. *Current Biology* *25*. 10.1016/j.cub.2015.08.030.
- 583 28. Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic Feature Encoding in Human
584 Superior Temporal Gyrus. *Science* *343*, 1006-1010. 10.1126/science.1245994.
- 585 29. Teoh, E.S., Ahmed, F., and Lalor, E.C. (2022). Attention Differentially Affects Acoustic and Phonetic
586 Feature Encoding in a Multispeaker Environment. *The Journal of Neuroscience* *42*, 682-691.
587 10.1523/jneurosci.1455-20.2021.
- 588 30. Lesenfans, D., Vanthornhout, J., Verschueren, E., and Francart, T. (2019). Data-driven spatial filtering
589 for improved measurement of cortical tracking of multiple representations of speech. *Journal of*
590 *Neural Engineering*. 10.1088/1741-2552/ab3c92.
- 591 31. Brodbeck, C., Hong, L.E., and Simon, J.Z. (2018). Rapid Transformation from Auditory to Linguistic
592 Representations of Continuous Speech. *Current Biology* *28*, 3976-3983.e3975.

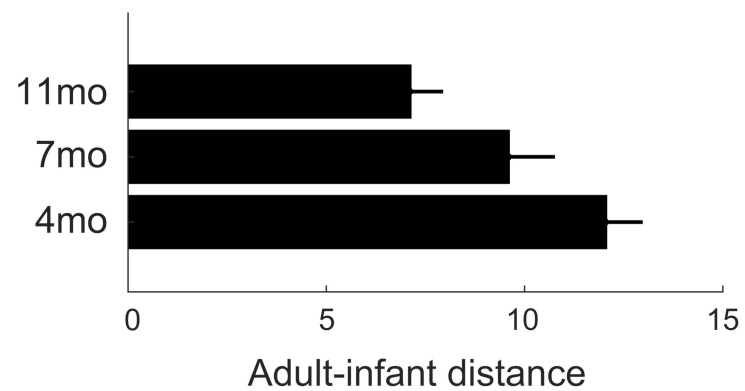
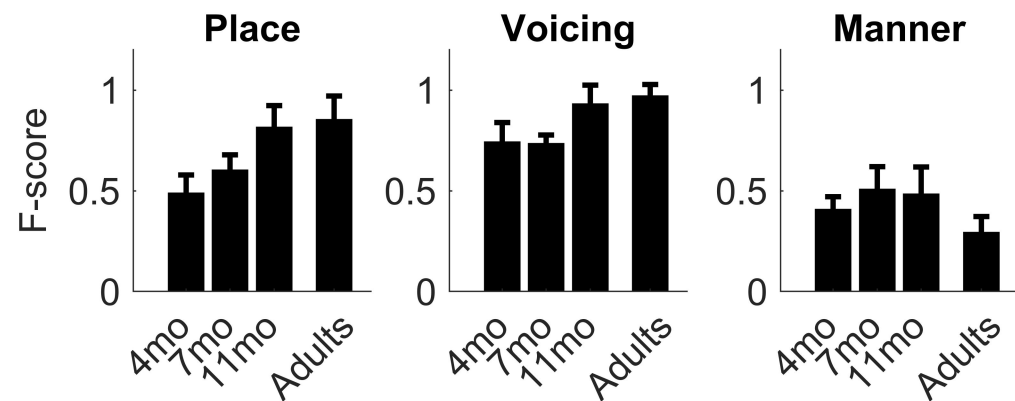
- 593 32. Di Liberto, G.M., Peter, V., Kalashnikova, M., Goswami, U., Burnham, D., and Lalor, E.C. (2018).
594 Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in
595 dyslexia. *NeuroImage NIMG-17-29*, 70-79. [10.1016/j.NEUROIMAGE.2018.03.072](https://doi.org/10.1016/j.NEUROIMAGE.2018.03.072).
- 596 33. Di Liberto, G.M., Nie, J., Yeaton, J., Khalighinejad, B., Shamma, S.A., and Mesgarani, N. (2021). Neural
597 representation of linguistic feature hierarchy reflects second-language proficiency. *NeuroImage* 227,
598 117586-117586. [10.1016/j.neuroimage.2020.117586](https://doi.org/10.1016/j.neuroimage.2020.117586).
- 599 34. Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (2008).
600 Phonetic learning as a pathway to language: New data and native language magnet theory expanded
601 (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society.
- 602 35. Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., and Lindblom, B. (1992). Linguistic experience
603 alters phonetic perception in infants by 6 months of age. *Science* 255, 606-608.
- 604 36. Dehaene-Lambertz, G., and Spelke, E.S. (2015). The infancy of the human brain. *Neuron* 88, 93-109.
- 605 37. Tsao, F.-M., Liu, H.-M., and Kuhl, P.K. (2004). Speech Perception in Infancy Predicts Language
606 Development in the Second Year of Life: A Longitudinal Study. *Child Development* 75, 1067-1084.
- 607 38. Kuhl, P.K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a
608 facilitation effect for native language phonetic perception between 6 and 12 months. *Dev Sci* 9, F13-
609 f21. [10.1111/j.1467-7687.2006.00468.x](https://doi.org/10.1111/j.1467-7687.2006.00468.x).
- 610 39. Eilers, R.E., Wilson, W.R., and Moore, J.M. (1977). Developmental changes in speech discrimination
611 in infants. *J Speech Hear Res* 20, 766-780. [10.1044/jshr.2004.766](https://doi.org/10.1044/jshr.2004.766).
- 612 40. Kuhl, P.K., Conboy, B.T., Padden, D., Nelson, T., and Pruitt, J. (2005). Early Speech Perception and
613 Later Language Development: Implications for the "Critical Period". *Language Learning and*
614 *Development* 1, 237-264. [10.1080/15475441.2005.9671948](https://doi.org/10.1080/15475441.2005.9671948).
- 615 41. Polka, L., Colantonio, C., and Sundara, M. (2001). A cross-language comparison of /d/-/th/
616 perception: evidence for a new developmental pattern. *J Acoust Soc Am* 109, 2190-2201.
617 [10.1121/1.1362689](https://doi.org/10.1121/1.1362689).
- 618 42. Di Liberto, G.M., and Lalor, E.C. (2017). Indexing cortical entrainment to natural speech at the
619 phonemic level: Methodological considerations for applied research. *Hearing Research* 348, 70-77.
620 [10.1016/j.heares.2017.02.015](https://doi.org/10.1016/j.heares.2017.02.015).
- 621 43. Di Liberto, G.M., Crosse, M.J., and Lalor, E.C. (2018). Cortical Measures of Phoneme-Level Speech
622 Encoding Correlate with the Perceived Clarity of Natural Speech. *Eneuro* 5, ENEURO.0084-0018.2018.
623 [10.1523/ENEURO.0084-18.2018](https://doi.org/10.1523/ENEURO.0084-18.2018).
- 624 44. Di Liberto, G.M., Wong, D., Melnik, G.A., and de Cheveigne, A. (2019). Low-frequency cortical
625 responses to natural speech reflect probabilistic phonotactics. *NeuroImage* 196, 237-247.
626 [10.1016/j.neuroimage.2019.04.037](https://doi.org/10.1016/j.neuroimage.2019.04.037).
- 627 45. de Cheveigné, A., Di Liberto, G.M., Arzounian, D., Wong, D.D.E., Hjortkjær, J., Fuglsang, S., and Parra,
628 L.C. (2019). Multiway canonical correlation analysis of brain data. *NeuroImage* 186, 728-740.
629 [10.1016/j.NEUROIMAGE.2018.11.026](https://doi.org/10.1016/j.NEUROIMAGE.2018.11.026).
- 630 46. Cooper, R.P., and Aslin, R.N. (1990). Preference for infant-directed speech in the first month after
631 birth. *Child Dev* 61, 1584-1595.
- 632 47. Gasparini, L., Langus, A., Tsuji, S., and Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in
633 infants' language discrimination abilities: A meta-analysis. *Cognition* 213, 104757.
634 <https://doi.org/10.1016/j.cognition.2021.104757>.
- 635 48. Jusczyk, P.W., and Aslin, R.N. (1995). Infants' Detection of the Sound Patterns of Words in Fluent
636 Speech. *Cognitive Psychology* 29, 1-23. <https://doi.org/10.1006/cogp.1995.1010>.
- 637 49. Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957). The discrimination of speech
638 sounds within and across phoneme boundaries. *J Exp Psychol* 54, 358-368. [10.1037/h0044417](https://doi.org/10.1037/h0044417).
- 639 50. Feldman, N.H., Goldwater, S., Dupoux, E., and Schatz, T. (2021). Do Infants Really Learn Phonetic
640 Categories? *Open Mind* 5, 113-131. [10.1162/opmi_a_00046](https://doi.org/10.1162/opmi_a_00046).
- 641 51. Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2010). Categorical
642 speech representation in human superior temporal gyrus. *Nat Neurosci* 13, 1428-1432.
643 <http://www.nature.com/neuro/journal/v13/n11/abs/nn.2641.html#supplementary-information>.

- 644 52. Huotilainen, M., Kujala, A., Hotakainen, M., Parkkonen, L., Taulu, S., Simola, J., Nenonen, J.,
645 Karjalainen, M., and Näätänen, R. (2005). Short-term memory functions of the human fetus recorded
646 with magnetoencephalography. *NeuroReport* *16*, 81-84.
- 647 53. Cheour, M., Alho, K., Čeponienė, R., Reinikainen, K., Sainio, K., Pohjavuori, M., Aaltonen, O., and
648 Näätänen, R. (1998). Maturation of mismatch negativity in infants. *International Journal of*
649 *Psychophysiology* *29*, 217-226. [https://doi.org/10.1016/S0167-8760\(98\)00017-8](https://doi.org/10.1016/S0167-8760(98)00017-8).
- 650 54. Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku,
651 P., Ilmoniemi, R.J., Luuk, A., et al. (1997). Language-specific phoneme representations revealed by
652 electric and magnetic brain responses. *Nature* *385*, 432-434. 10.1038/385432a0.
- 653 55. Cheour, M., Martynova, O., Näätänen, R., Erkkola, R., Sillanpää, M., Kero, P., Raz, A., Kaipio, M.-L.,
654 Hiltunen, J., and Aaltonen, O. (2002). Speech sounds learned by sleeping newborns. *Nature* *415*, 599-
655 600.
- 656 56. Choi, D., Dehaene-Lambertz, G., Peña, M., and Werker, J.F. (2021). Neural indicators of articulator-
657 specific sensorimotor influences on infant speech perception. *Proceedings of the National Academy*
658 *of Sciences* *118*, e2025043118. doi:10.1073/pnas.2025043118.
- 659 57. Gennari, G., Marti, S., Palu, M., Fló, A., and Dehaene-Lambertz, G. (2021). Orthogonal neural codes
660 for speech in the infant brain. *Proceedings of the National Academy of Sciences* *118*, e2020410118.
661 doi:10.1073/pnas.2020410118.
- 662 58. Leong, V., and Goswami, U. (2015). Acoustic-Emergent Phonology in the Amplitude Envelope of
663 Child-Directed Speech. *PLOS ONE* *10*, e0144411. 10.1371/journal.pone.0144411.
- 664 59. Leong, V., Kalashnikova, M., Burnham, D., and Goswami, U. (2017). The Temporal Modulation
665 Structure of Infant-Directed Speech. *Open Mind* *1*, 78-90. 10.1162/OPMI_a_00008.
- 666 60. Daube, C., Ince, R.A.A., and Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based
667 Predictions of Cortical Responses to Speech. *Current Biology* *29*, 1924-1937.e1929.
668 10.1016/J.CUB.2019.04.067.
- 669 61. Gillis, M., Vanthornhout, J., Simon, J.Z., Francart, T., and Brodbeck, C. (2021). Neural Markers of
670 Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling
671 the Speech Acoustics. *The Journal of Neuroscience* *41*, 10316. 10.1523/JNEUROSCI.0812-21.2021.
- 672 62. Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., and Lalor, E.C. (2018).
673 Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural,
674 Narrative Speech. *Current Biology*. 10.1016/j.cub.2018.01.080.
- 675 63. Decruy, L., Vanthornhout, J., and Francart, T. (2020). Hearing impairment is associated with enhanced
676 neural tracking of the speech envelope. *Hearing Research* *393*, 107961-107961.
677 10.1016/j.heares.2020.107961.
- 678 64. Brodbeck, C., Presacco, A., Anderson, S., and Simon, J.Z. (2018). Over-representation of speech in
679 older adults originates from early response in higher order auditory cortex. 2018/9//. (S. Hirzel Verlag
680 GmbH), pp. 774-777.
- 681 65. Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., and Francart, T. (2018). Speech Intelligibility
682 Predicted from Neural Entrainment of the Speech Envelope. *Journal of the Association for Research*
683 *in Otolaryngology* *19*, 181-191. 10.1007/s10162-018-0654-z.
- 684 66. Crosse, M.J., Di Liberto, G.M., and Lalor, E.C. (2016). Eye can hear clearly now: Inverse effectiveness
685 in natural audiovisual speech processing relies on long-term crossmodal temporal integration.
686 *Journal of Neuroscience* *36*. 10.1523/JNEUROSCI.1396-16.2016.
- 687 67. Menn, K.H., Ward, E.K., Braukmann, R., van den Boomen, C., Buitelaar, J., Hunnius, S., and Snijders,
688 T.M. (2022). Neural Tracking in Infancy Predicts Language Development in Children With and Without
689 Family History of Autism. *Neurobiology of Language* *3*, 495-514. 10.1162/nol_a_00074.
- 690 68. Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG
691 dynamics including independent component analysis. *J Neurosci Methods* *134*, 9-21.
692 10.1016/j.jneumeth.2003.10.009.
- 693 69. Greenwood, D.D. (1961). Auditory Masking and the Critical Band. *The Journal of the Acoustical*
694 *Society of America* *33*, 484-502. doi:<http://dx.doi.org/10.1121/1.1908699>.
- 695 70. Chomsky, N., and Halle, M. (1968). The sound pattern of English.

- 696 71. Ladefoged, P., and Johnson, K. (2014). A course in phonetics.
697 72. Boersma, P., and Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1. 05)[Computer
698 program]. Retrieved May 1, 2009.
699 73. Ding, N., Chatterjee, M., and Simon, J.Z. (2014). Robust cortical entrainment to the speech envelope
700 relies on the spectro-temporal fine structure. *NeuroImage* 88, 41-46.
701 74. Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika* 28, 321-321.
702 10.2307/2333955.
703 75. Harris, P.A., Taylor, R., Minor, B.L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G.,
704 Delacqua, F., Kirby, J., and Duda, S.N. (2019). The REDCap consortium: Building an international
705 community of software platform partners. *J Biomed Inform* 95, 103208. 10.1016/j.jbi.2019.103208.
706 76. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J.G. (2009). Research electronic
707 data capture (REDCap)--a metadata-driven methodology and workflow process for providing
708 translational research informatics support. *J Biomed Inform* 42, 377-381. 10.1016/j.jbi.2008.08.010.
709

A**B****C****D**

A**Hypotheses****B****0 - 400ms****100 - 500ms****C****D**

A**B****C**