Ghinst, Marc Vander; Bourguignon, Mathieu; Niesen, Maxime; Wens, Vincent; Hassid, Sergio; Choufani, Georges; Jousmäki, Veikko; Hari, Riitta; Goldman, Serge; De Tiège, Xavier

Cortical tracking of speech-in-noise develops from childhood to adulthood

Published in:
JOURNAL OF NEUROSCIENCE

DOI:
10.1523/JNEUROSCI.1732-18.2019

Published: 10/04/2019

Document Version
Peer reviewed version

Please cite the original version:

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.
Research Articles: Behavioral/Cognitive

Cortical tracking of speech-in-noise develops from childhood to adulthood

Marc Vander Ghinst1,2, Mathieu Bourguignon1,3,4, Maxime Niesen1,2, Vincent Wens1,5, Sergio Hassid2, Georges Choufani2, Veikko Jousmäki6,7, Riitta Hari6, Serge Goldman1,5 and Xavier De Tiège1,5

1Laboratoire de Cartographie fonctionnelle du Cerveau, UNI — ULB Neuroscience Institute, Université libre de Bruxelles (ULB), 1070 Brussels, Belgium.
2Service d’ORL et de chirurgie cervico-faciale, ULB-Hôpital Erasme, Université libre de Bruxelles (ULB), 1070 Brussels, Belgium.
3Laboratoire Cognition Langage et Développement, UNI — ULB Neuroscience Institute, Université libre de Bruxelles (ULB), Brussels, Belgium.
4Basque Center on Cognition, Brain and Language (BCBL), Donostia/San Sebastian, Spain.
5Department of Functional Neuroimaging, Service of Nuclear Medicine, CUB Hôpital Erasme, Université libre de Bruxelles (ULB), Brussels, Belgium.
6Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, PO BOX 15100, FI-00076-AALTO, Espoo, Finland.
7Cognitive Neuroimaging Centre, Lee Kong Chian School of Medicine, 59 Nanyang Drive, Nanyang Technological University, Singapore 639591.
8Department of Art, Aalto University School of Arts, Design, and Architecture., PO Box 31000, 00076 AALTO, Helsinki, Finland.

https://doi.org/10.1523/JNEUROSCI.1732-18.2019

Received: 10 July 2018
Revised: 8 January 2019
Accepted: 12 January 2019
Published: 11 February 2019


Conflict of Interest: The authors declare no competing financial interests.

Marc Vander Ghinst and Maxime Niesen were supported by a research grant from the Fonds Erasme (Brussels, Belgium). Mathieu Bourguignon was supported by the program Attract of Innoviris (grant 2015-BB2B-10), by the Spanish Ministry of Economy and Competitiveness (grant PSI2016-77175-P), and by the Marie Skłodowska-Curie Action of the European Commission (grant 743562). Veikko Jousmäki was supported by a research grant from the Institut d’Encouragement de la Recherche Scientifique et de l’Innovation de Bruxelles (“Brains back to Brussels”, Brussels, Belgium). Xavier De Tiège is Post-doctorate Clinical Master Specialist at the FRS-FNRS.

Corresponding author: Marc Vander Ghinst, Laboratoire de Cartographie fonctionnelle du Cerveau, UNI — ULB Neuroscience Institute, Université libre de Bruxelles (ULB), 808 Lennik Street, 1070 Brussels, Belgium. Telephone: +32.2.555.31.11, Fax: +32.2.555.47.01, E-mail: Marc.Vander.Ghinst@erasme.ulb.ac.be

Cite as: J. Neurosci 2019; 10.1523/JNEUROSCI.1732-18.2019

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.
Cortical tracking of speech-in-noise develops from childhood to adulthood

Abbreviated Title: Cortical processing of speech-in-noise in children

Marc Vander Ghinst\textsuperscript{1,2,*}, Mathieu Bourguignon\textsuperscript{1,3,4,*}, Maxime Niesen\textsuperscript{1,2}, Vincent Wens\textsuperscript{1,5}, Sergio Hassid\textsuperscript{2}, Georges Choufani\textsuperscript{2}, Veikko Jousmäki\textsuperscript{6,7}, Riitta Hari\textsuperscript{8}, Serge Goldman\textsuperscript{1,5}, and Xavier De Tiège\textsuperscript{1,5}.

\textsuperscript{1} Laboratoire de Cartographie fonctionnelle du Cerveau, UNI – ULB Neuroscience Institute, Université libre de Bruxelles (ULB), 1070 Brussels, Belgium.
\textsuperscript{2} Service d’ORL et de chirurgie cervico-faciale, ULB-Hôpital Erasme, Université libre de Bruxelles (ULB), 1070 Brussels, Belgium.
\textsuperscript{3} Laboratoire Cognition Langage et Développement, UNI – ULB Neuroscience Institute, Université libre de Bruxelles (ULB), Brussels, Belgium.
\textsuperscript{4} Basque Center on Cognition, Brain and Language (BCBL), Donostia/San Sebastian, Spain.
\textsuperscript{5} Department of Functional Neuroimaging, Service of Nuclear Medicine, CUB Hôpital Erasme, Université libre de Bruxelles (ULB), Brussels, Belgium.
\textsuperscript{6} Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, PO BOX 15100, FI-00076-AALTO, Espoo, Finland.
\textsuperscript{7} Cognitive Neuroimaging Centre, Lee Kong Chian School of Medicine, 59 Nanyang Drive, Nanyang Technological University, Singapore 636921.
\textsuperscript{8} Department of Art, Aalto University School of Arts, Design, and Architecture, PO Box 31000, 00076 AALTO, Helsinki, Finland.

*These authors equally contributed to this work.

Corresponding author: Marc Vander Ghinst, Laboratoire de Cartographie fonctionnelle du Cerveau, UNI – ULB Neuroscience Institute, Université libre de Bruxelles (ULB), 808 Lennik Street, 1070 Brussels, Belgium. Telephone: +32.2.555.31.11, Fax: +32.2.555.47.01, E-mail: Marc.Vander.Ghinst@erasme.ulb.ac.be.

Number of figures: 7, number of table: 1, number of words in Abstract: 250, number of words in Introduction: 497, number of words in Discussion: 1500.
**Conflict of interest**

The authors have no conflict of interest to declare.

**Acknowledgments**

Marc Vander Ghinst and Maxime Niesen were supported by a research grant from the Fonds Erasme (Brussels, Belgium). Mathieu Bourguignon was supported by the program Attract of Innoviris (grant 2015-BB2B-10), by the Spanish Ministry of Economy and Competitiveness (grant PSI2016-77175-P), and by the Marie Skłodowska-Curie Action of the European Commission (grant 743562). Veikko Jousmäki was supported by a research grant from the Institut d'Encouragement de la Recherche Scientifique et de l'Innovation de Bruxelles ("Brains back to Brussels", Brussels, Belgium). Xavier De Tiège is Post-doctorate Clinical Master Specialist at the FRS-FNRS.

This study and the MEG project at the CUB Hôpital Erasme was financially supported by the Fonds Erasme (research convention “Les Voies du Savoir”, Fonds Erasme, Brussels, Belgium).

We thank Nicolas Grimault and Fabien Perrin (Centre de Recherche en Neuroscience, Lyon, France) for providing us their entire audio material. We also thank Brice Marty for providing his help for data acquisition.
Abstract

In multitalker backgrounds, the auditory cortex of adult humans tracks the attended speech stream rather than the global auditory scene. Still, it is unknown whether such preferential tracking also occurs in children whose speech-in-noise (SiN) abilities are typically lower compared with adults.

We used magnetoencephalography (MEG) to investigate the frequency-specific cortical tracking of different elements of a cocktail-party auditory scene in twenty children (6–9 years; 8 females) and twenty adults (21–40 years; 10 females). During MEG recordings, subjects attended to 4 different 5-min stories, mixed with different levels of multitalker background at four signal-to-noise ratios (SNRs: noiseless, +5, 0, and –5 dB). Coherence analysis quantified the coupling between the time courses of the MEG activity and attended speech stream, multitalker background or global auditory scene, respectively.

In adults, statistically significant coherence was observed between MEG signals originating from the auditory system and the attended stream at < 1 Hz, 1–4 Hz, and 4–8 Hz in all SNR conditions. Children displayed similar coupling at < 1 Hz and 1–4 Hz, but increasing noise impaired the coupling more strongly than in adults. Also, children displayed drastically lower coherence at 4–8 Hz in all SNR conditions.

These results suggest that children’s difficulties to understand speech in noisy conditions are related to an immature selective cortical tracking of the attended speech streams. Our results also provide unprecedented evidence for an acquired cortical tracking of speech at syllable-rate and argue for a progressive development of SiN abilities in humans.
Significance statement:

Behaviorally, children are worse than adults at understanding speech-in-noise. Here, neuromagnetic signals were recorded while healthy adults and typically developing 6–9-year-old children attended to a speech stream embedded in a multitalker background noise with varying intensity. Results demonstrate that auditory cortices of both children and adults selectively track the attended speaker’s voice rather than the global acoustic input at phrasal and word rates. However, increments of noise compromised the tracking significantly more in children than in adults. Unexpectedly, children displayed limited tracking of both the attended voice and the global acoustic input at the 4–8-Hz syllable rhythm. Thus, both speech-in-noise abilities and cortical tracking of speech syllable repetition rate seem to mature later in adolescence.
Introduction

Children often grow up and learn in noisy surroundings. Clamorous classrooms, rowdy playgrounds, and domestic sound disturbances indeed constitute adverse auditory scenes for a still immature auditory system.

Interestingly, speech-in-noise (SiN) perception in children appears strenuous but improves during late childhood (≥ 10 years) due to maturation of the auditory system and attentional abilities (Elliott, 1979; Moore et al., 2010; Sanes and Woolley, 2011; Thompson et al., 2017). Still, the neurophysiological mechanisms accounting for the improvement in SiN perception observed from childhood to adulthood are unsettled. It has been hypothesized that, in adverse auditory scenes, children’s auditory system would actually lack the capacity to segregate the attended auditory stream from the unattended noisy background (Sanes and Woolley, 2011). However, no study has so far demonstrated such phenomena in children.

Accumulating evidence shows that adults’ auditory system tracks the attended speech stream rather than the global auditory scene in a multitalker background (Mesgarani and Chang, 2012). In such adverse auditory background, the auditory system entrains to the slow amplitude modulations (i.e., the temporal envelope) of the attended speaker’s voice rather than to modulations of the global auditory scene (Ding and Simon, 2012a; Zion Golumbic et al., 2013; Vander Ghinst et al., 2016). This coupling typically occurs at frequencies below 10 Hz and declines with increasing noise level. Given that this frequency range matches with prosodic stress/phasisal/sentential (< 1 Hz), word (1–4 Hz), and syllable (4–8 Hz) repetition rates, the corresponding cortical tracking of speech has been hypothesized to subserve the chunking of the continuous verbal flow into relevant segments used for further speech recognition, up to a certain noise level (Ding and Simon, 2012a; Giraud and Poeppel, 2012; Ding and Simon, 2013a; Vander Ghinst et al., 2016; Keitel et al., 2018).
How children’s auditory system tracks connected speech has remained largely unknown. Until now, cortical tracking of speech in children has only involved dyslexic children older than 11 years, showing that they have impaired tracking below 2 Hz compared with age-matched healthy control subjects (Molinaro et al., 2016; Power et al., 2016). However, to the best of our knowledge, no study has so far investigated how this low-frequency cortical tracking might differ between typically developed children and adults, and whether noise differentially corrupts the coupling in these two populations. We therefore specifically test the hypothesis that children’s poor SiN perception abilities (Elliott, 1979; Berman and Friedman, 1995) are related to inaccurate low-frequency cortical tracking of the attended speech stream in a noisy background.

To test this hypothesis, children (6–9 years) and adults (21–40 years) with normal SiN perception listened to speech recordings mixed with a cocktail party noise at different intensities in an ecological connected speech listening paradigm. Based on magnetoencephalographic (MEG) recordings, we quantified the cortical tracking of the different elements of the auditory scene: (i) Attended stream (i.e., reader’s voice only), (ii) unattended Multitalker background only, and (iii) Global scene (i.e., the combination of the Attended stream and the unattended Multitalker background).
Methods

The methods used for MEG data acquisition, preprocessing, and analyses are derived from previous studies (Bourguignon et al., 2013; Vander Ghinst et al., 2016).

Participants

Twenty native French-speaking healthy children (mean age 8 yrs, range 6–9 yrs; 8 females and 12 males) and twenty native French-speaking healthy adults (mean age 30 yrs, range 21–40 yrs; 10 females and 10 males) without any history of neuropsychiatric or otologic disorder participated in this study. All subjects had normal hearing according to pure tone audiometry (i.e., normal hearing thresholds (between 0–20 dB HL) for 250, 500, 1000, 2000, 4000, and 8000 Hz) and normal otomicroscopy. Subjects' auditory perception was assessed with three separate subtests of a validated and standardized French language central auditory battery: 1) a dichotic test, 2) a speech audiometry, and 3) a SiN audiometry (Demanez et al., 2003). In the two later tests, 30 monosyllabic words were presented with (3) or without (2) noise in a predetermined counterbalanced order, so that every word is presented once in silence, and once in noise. A score was then obtained, corresponding to the number of words correctly repeated with and without noise. According to the tests (1–3), all subjects had normal dichotic perception, speech and SiN perception for their age (Demanez et al., 2003). Children and adults were all right-handed according to the Edinburgh handedness inventory (Oldfield, 1971). The study had prior approval by the ULB-Hôpital Erasme Ethics Committee. Participants gave written informed consent before participation.

Experimental paradigm

During MEG recordings, the subjects sat comfortably in the MEG chair with the arms resting on a table positioned in front of them. They underwent four listening conditions and
one rest condition, each lasting five minutes. The order of the five conditions was randomized for each subject.

Subjects were told before the task that questions on the content of the story would be asked after each listening condition. Children were given a clue about the content of the text they were about to listen to ensure that they selected the attended auditory stream straightaway (e.g., “you are going to listen to the story of two little princes”). During the listening conditions, subjects listened to four different stories in French recorded by different native adult French-speakers. The recordings were randomly selected from a set of four stories (readers’ sex ratio: 1/1) obtained from a French audiobook database (http://www.litteratureaudio.com) after written authorization from the readers. Children and adults listened to different stories adapted to their age to maximize their implication in the task, and their comprehension of the stories. This approach was particularly important for stories used in children as it has been previously demonstrated that reading stories aloud from books exposes children to a linguistic and cognitive complexity typically not found in child-directed or adult-directed speech (Massaro, 2017). Special care was therefore taken to select stories with comprehensible vocabulary and content. Phrasal, word, and syllable rates, assessed as the number of phrases, words, or syllables divided by the corrected duration of the audio recording, were comparable in children (mean phrasal, word and syllable rates across different stories 0.45 Hz, 3.6 Hz, and 5.54 Hz) and in adults (0.49 Hz, 3.39 Hz, and 5.56 Hz, respectively). For phrases, the corrected duration was (trivially) the total duration of the audio recording. For words and syllables, the corrected duration was the total time during which the speaker was actually speaking, that is the total duration of the audio recording (here 5 min) minus the sum of all silent periods when the speech amplitude was below a tenth of the mean amplitude for at least 10 ms. A specific speech signal-to-noise ratio (SNR; where signal was attended reader’s voice, and noise was the multi-talker background) was randomly assigned to
each story: a noiseless condition, and 3 SiN conditions (with SNRs of +5 dB, 0 dB, and –5 dB), leading to four different SNR conditions. This randomization procedure prevented any systematic association between stories and SNRs. The noise (Fonds sonores v-1.0, Perrin & Grimault 2005) was a continuous cocktail party noise obtained by mixing the voices of six native French speakers talking simultaneously in French (three females and three males). This configuration of cocktail-party noise was selected because it accounts for both energetic and informational masking at phonetic and lexical level (Simpson and Cooke, 2005; Hoen et al., 2007). Sound recordings were played using VLC media player (VideoLAN Project, GNU General Public License) running on a MacBook Pro (Apple Computer, Inc., Cupertino, CA). Sound signals were transmitted to a MEG-compatible 60 × 60 cm² high-quality flat panel loudspeaker (Panphonics SSH sound shower, Panphonics Oy, Espoo, Finland) placed 2.4 m away in front of the subjects. The average sound intensity was set to 60 dB as assessed by a sound level meter (Sphynx Audio System, Belgium). Subjects were asked to attend to the reader’s voice and to gaze at a fixation point on the wall of the magnetically shielded room facing them. During the Rest condition, subjects were instructed to relax, not to move, and to gaze at the same fixation point. At the end of each listening condition, subjects were asked to score the intelligibility of the attended reader’s voice on a visual analog scale (VAS) ranging from 0 to 10 (0 = totally unintelligible; 10 = perfectly intelligible) and were also asked 16 (adults) or 8 (children) yes/no forced-choice questions exploring the salience and explicitness of the heard story by analogy with what is required for the clinical diagnosis of text comprehension deficits (Ferstl et al., 2005).

**Data acquisition**

Cortical neuromagnetic signals were recorded at ULB-Hôpital Erasme using a whole-scalp-covering 306-channel MEG device (for 15 adults and 12 children Elekta Neuromag...
Vectorview by Elekta Oy, Helsinki, Finland, and otherwise Elekta Neuromag Triux by MEGIN, Helsinki, Finland) installed in a light-weight magnetically shielded room (Maxshield, MEGIN, Helsinki, Finland), the characteristics of which being described elsewhere (De Tiège et al., 2008; Carrette et al., 2011). The MEG device has 102 sensor chipsets, each comprising one magnetometer and two orthogonal planar gradiometers. MEG signals were bandpass-filtered through 0.1–330 Hz and sampled at 1 kHz. Four head-tracking coils monitored subjects’ head position inside the MEG helmet. The locations of the coils and at least 150 head-surface (on scalp, nose, and face) points with respect to anatomical fiducials were digitized with an electromagnetic tracker (Fastrack, Polhemus, Colchester, VT). Electro-oculogram (EOG), electrocardiogram (ECG), and the audio signals presented to the subjects were recorded simultaneously with MEG signals (passband 0.1–330 Hz for EOG and ECG, and low-pass at 330 Hz for audio signals). The recorded audio signals were used for synchronization between MEG and the transmitted audio signals, the latter being bandpass-filtered at 50–22000 Hz and sampled at 44.1 kHz. High-resolution 3D-T1 cerebral magnetic resonance images (MRI) were acquired on a 1.5 T MRI (Intera, Philips, The Netherlands).

**Data preprocessing**

Continuous MEG data were first preprocessed off-line using the temporal extension of the signal-space-separation method (correlation limit, 0.9; segment length, 20 s) to suppress external inferences and correct for head movements (Taulu et al., 2005; Taulu and Simola, 2006). For the subsequent coherence analyses used to quantify the cortical tracking of speech, continuous MEG and audio signals were split into 2048-ms epochs with 1638-ms epoch overlap, leading to a frequency resolution of ~0.5 Hz (Bortel and Sovka, 2007). MEG epochs exceeding 3 pT (magnetometers) or 0.7 pT/cm (gradiometers) were excluded from further analysis to avoid contamination of the data by eye movement artifacts, muscle activity, or...
artifacts in the MEG sensors. The number of artifact-free epochs was 695 ± 66 (mean ± SD across subjects and conditions) in the adult group and 621 ± 93 in the children group.

A two-way repeated-measures ANOVA (factors: age group and condition; dependent variable: number of epochs used to compute coherence) revealed a significant effect of group on the number of epochs ($F_{1,114} = 4.42$, $p = 0.042$) but no effect of SNR ($F_{3,114} = 0.03$, $p = 0.99$) or interaction ($F_{3,114} = 1.46$, $p = 0.23$). To avoid a possible methodological bias in our results due to differences between age groups in the accuracy of speech tracking estimation, we threw away epochs in adults’ data so as to equalize the number of epochs in both groups.

**Coherence analyses in sensor space**

For each listening condition, synchronization between the temporal envelope of wide-band (50–22000 Hz) audio signals and artifact-free MEG epochs (2048-ms-long) was assessed with coherence analysis in sensor space at frequencies in which speech temporal envelope is critical for speech comprehension (i.e., 0.1–20 Hz) (Drullman et al., 1994). Coherence is an extension of Pearson correlation coefficient to the frequency domain. It quantifies the degree of coupling between two signals (say $x(t)$ and $y(t)$), providing a number between 0 (no linear dependency) and 1 (perfect linear dependency) for each frequency bin (Halliday et al., 1995). Coherence was computed as follows:

$$\text{Coh}_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f)P_{yy}(f)},$$

where $P_{xx}(f) = \sum_k |\hat{x}_k(f)|^2$, $P_{yy}(f) = \sum_k |\hat{y}_k(f)|^2$, and $P_{xy}(f) = \sum_k |\hat{x}_k(f)\hat{y}^*_k(f)|$, and

where $\hat{x}_k(f)$ (respectively $\hat{y}_k(f)$) is the Fourier coefficient of the $k^{th}$ epoch of signal $x(t)$ (respectively $y(t)$) at frequency bin $f$, and $^*$ denotes the complex conjugate. In our
application, coherence analysis provides a quantitative assessment of the cortical tracking of speech.

For the three SiN conditions (+5 dB, 0 dB, and –5 dB), coherence was separately computed between MEG signals and three acoustic elements of the auditory scene: 1) the Global scene (Attended stream + Multitalker background), leading to $Coh_{\text{global}}$, 2) the Attended stream only (i.e., the reader’s voice), leading to $Coh_{\text{att}}$, and 3) the Multitalker background only, leading to $Coh_{\text{bg}}$. Sensor-level coherence maps were obtained using gradiometer signals only, and signals from gradiometer pairs were combined as done in Bourguignon et al. (2015).

Previous studies have demonstrated statistically significant coupling between acoustic and brain signals at frequencies corresponding to phrases, words, and syllables (Ding and Simon, 2012b; Bourguignon et al., 2013; Peelle et al., 2013; Clumeck et al., 2014; Koskinen and Seppä, 2014; Vander Ghinst et al., 2016, Keitel et al., 2018). Accordingly, sensor-level coherence maps were produced separately for phrases (frequency bin corresponding to 0.5 Hz), words (average across the frequency bins falling in 1–4 Hz) and syllables (4–8 Hz). Note that coherence at the frequency bin corresponding to 0.5 Hz actually reflects coupling in a frequency range around 0.5 Hz, with a sensitivity profile proportional to the Fourier transform of a boxcar function: $\text{sinc}(\pi(f-0.5) / 0.5)$. The < 1 Hz, 1–4 Hz, and 4–8 Hz frequency ranges are henceforth referred to as frequency bands of interest.

Coherence analyses in source space

Individual MRIs were first segmented using Freesurfer software (Martinos Center for Biomedical Imaging, Boston, MA, RRID:SCR_001847; (Reuter et al., 2012)). MEG and segmented MRI coordinate systems were then coregistered using the three anatomical fiducial points for initial estimation and the head-surface points to manually refine the surface.
coregistration. The MEG forward model was computed for triplets of orthogonal current
dipoles, placed on a homogenous 5-mm-grid source space covering the whole brain, using
MNE suite (Martinos Centre for Biomedical Imaging, Boston, MA, RRID:SCR_005972;
(Gramfort et al., 2014)). The forward model was then reduced to its two first principal
components. This procedure is justified by the insensitivity of MEG to currents radial to the
skull, and hence, this dimension reduction leads to considering only the tangential sources. As
a preliminary step, to simultaneously combine data from planar gradiometers and
magnetometers for source estimation, sensor signals (and the corresponding forward-model
coefficients) were normalized by their noise root mean square, estimated from the Rest data
filtered through 1–195 Hz. Coherence maps obtained for each subject, listening condition
(Noiseless, +5 dB, 0 dB, and –5 dB), audio signal (Global scene, Attended stream, Multitalker
Background) and frequency bands of interest (< 1 Hz, 1–4 Hz, and 4–8 Hz) were finally
produced using the Dynamic Imaging of Coherent Sources approach (Gross et al., 2001) with
Minimum-Norm Estimates inverse solution (Dale and Sereno, 1993). Noise covariance was
estimated from the Rest data filtered through 1–195 Hz and the regularization parameter was
fixed in terms of MEG sensor noise level as done by Hämäläinen et al. (2010).

Group-level analyses in source space
A non-linear transformation from individual MRIs to the standard Montreal
Neurological Institute (MNI) brain was first computed using the spatial normalization
algorithm implemented in Statistical Parametric Mapping (SPM8, Wellcome Department of
Cognitive Neurology, London, UK, RRID:SCR_007037; (Ashburner et al., 1997; Ashburner
and Friston, 1999)) and then applied to individual MRIs and every coherence map. The adult
MNI template was used in both children and adults despite the fact that spatial normalization
may fail for brains of small size when using an adult template (Reiss et al., 1996). However,
this risk is negligible for the population studied here. Indeed, the brain volume does not change substantially from the age of 5 years to adulthood (Reiss et al., 1996). This assumption has been confirmed by a study that specifically addressed this question in children aged above 6 years (Muzik et al., 2000).

To produce coherence maps at the group level, we computed across subjects the generalized \( f \)-mean of normalized maps, according to \( f(\xi) = \arctanh(\sqrt{\xi}) \), namely, the Fisher \( z \)-transform of the square-root. This procedure transforms the noise on the coherence estimate into an approximately normally distributed noise (Rosenberg et al., 1989). Thus, the computed coherence is an unbiased estimate of the mean coherence at the group-level. In addition, this averaging procedure avoids an over-contribution of subjects characterized by high coherence values to the group analysis (Bourguignon et al., 2012). The resulting subject- and group-level coherence maps are henceforth referred to as the audio maps.

**Experimental design and statistical analyses**

Sample size was based on a previous study from our group with a similar design, which included 20 healthy adults (Vander Ghinst et al., 2016). Accordingly, we set the sample size to 20 per age group.

*Comparison of SiN perception in adults vs. children*

Children’ and adults’ capacities to understand speech and SiN—as measured with speech and SiN audiometry—were compared with a \( t \)-test.

*Effect of SNR on the comprehension and the intelligibility of the Attended stream.*

A two-way repeated-measures ANOVA was used to assess the effects of the multitalker background noise level (within-subject factor; *Noiseless*, +5 dB, 0 dB, and −5 dB) and of the
age group (between-subjects factor; adults, children) on the comprehension scores and intelligibility ratings separately. The distribution of the residues of the ANOVAs was then tested for normality using the Lilliefors test (Lilliefors, 1967).

Of note, we acknowledge that the interpretability of these analyses could be limited for two reasons. First, adults and children listened to different texts and had to answer questions where the difficulty was adapted to their age. Second, the intelligibility ratings by children and adults may differ also due to differences in the visual analogue scales: explicit visual support was provided for the children (more or less happy faces) to facilitate the evaluation.

Significance of individual subjects’ coherence values

The statistical significance of individual subjects’ coherence values (for each listening condition, audio signal, and frequency band of interest) was assessed with surrogate-data-based maximum statistics. This statistical assessment was performed on sensor-space coherence values, and it tested the null hypothesis that the brain does not track audio signals more than other plausible unrelated (surrogate) signals. This method was chosen because it intrinsically deals with the issue of multiple-comparisons across sensors, and as it takes into account the temporal auto-correlation within signals. For each subject, 1000 surrogate sensor-level coherence maps were computed as done for genuine coherence maps but with audio signals replaced by Fourier transform surrogate audio signals (Faes et al., 2004). The maximum coherence value across all sensors was extracted for each surrogate simulation, and the 95th percentile of this distribution of maximum coherence values yielded the significance threshold at $p < 0.05$.

Significance of group-level coherence values
The statistical significance of coherence values in group-level *audio* maps was assessed for each hemisphere separately with a non-parametric permutation test (Nichols and Holmes, 2002). In practice, subject- and group-level *Rest* coherence maps were computed as done for the *audio* maps, but with MEG signals in listening conditions replaced by *Rest* MEG signals and sound signals unchanged. Group-level difference maps were obtained by subtracting *f*-transformed *audio* and *Rest* group-level coherence maps. Under the null hypothesis that coherence maps are the same whatever the experimental condition, the labeling *audio* and *Rest* are exchangeable at the subject-level prior to group-level difference map computation (Nichols and Holmes, 2002). To reject this hypothesis and to compute a threshold of statistical significance for the correctly labeled difference map for each hemisphere separately, the permutation distribution of the maximum of the difference map’s absolute value in each hemisphere was computed for 10,000 permutations. The test assigned a *p* value to each voxel in the group-level *audio* map, equal to the proportion of surrogate values exceeding the corresponding voxel’s difference value (Nichols and Holmes, 2002).

We further identified the coordinates of local maxima in group-level coherence maps. Such local coherence maxima are sets of contiguous voxels displaying higher coherence values than all neighbouring voxels. We only report statistically significant local coherence maxima, disregarding the extents of these clusters. Indeed, cluster extent is hardly interpretable in view of the inherent smoothness of MEG source reconstruction (Hämäläinen et al., 1994; Wens et al., 2015; Bourguignon et al., 2017).

**Cortical processing of the auditory scene in SiN conditions**

To identify cortical areas wherein activity reflects more the *Attended stream* than the *Global scene*, we compared *Coh*\(_{att}\) to *Coh*\(_{global}\) maps using the same non-parametric
permutation test described above, but with the labels Global scene and Attended stream instead of audio and Rest, leading to the Coh\textsubscript{att} – Coh\textsubscript{global} difference maps.

Comparison of the tracking in adults vs. children

To identify cortical areas showing stronger tracking in adults than children (and vice versa), we compared the corresponding coherence maps using the above described permutation test, but with the labels Coh\textsubscript{att,adults} and Coh\textsubscript{att,children} instead of Coh\textsubscript{att} and Coh\textsubscript{global}, leading to the Coh\textsubscript{att,adults} – Coh\textsubscript{att,children} difference maps.

Effect of the SNR, age group, and hemispheric lateralization on cortical tracking of speech in noise

In this between-subject design, we used a 3-way repeated-measures ANOVA to compare cortical tracking of speech between n = 20 children and n = 20 adults with additional factors of hemisphere (left vs. right) and four different SNR conditions (Noiseless, +5 dB, 0 dB, and −5 dB). The dependent variable was the maximal Coh\textsubscript{att} value within a sphere of 10 mm radius around the maximum of the group-level difference map in each hemisphere. As the Lilliefors test revealed that the distribution of the residues of the ANOVAs deviated statistically significantly from the normal distribution for 1–4 Hz and 4–8 Hz Coh\textsubscript{att} values (p \textless 0.05), we repeated the ANOVAs on the coherence values transformed with the transformation used to average source coherence maps (f(·) = arctanh(√·)). After such transformation, the residues did not deviate significantly from a normal distribution (p \textgreater 0.05). Therefore, we report on only the results of the later ANOVAs. Post-hoc comparisons were performed with pairwise t-tests.
Based on our results in the Noiseless condition, we conducted an additional analysis by computing Pearson correlation between children’s age and their maximum coherence values, separately for both hemispheres in the three frequency bands of interest.
Results

In this study, children and adults were listening to connected speech embedded in a multitalker background with different SNR conditions. Ensuing cortical tracking of speech (i.e., the coupling between brain activity and audio signals) was quantified with coherence analysis. The specific aim was to compare this tracking between children and adults.

Comparison of SiN perception in adults vs. children

Speech perception in silence did not differ ($t_{38} = 1.27; p = 0.211$) between children (28.35 ± 0.88; mean ± SD) and adults (28.7 ± 0.86). SiN perception quantified with SiN audiometry was significantly ($t_{38} = 3.35; p = 0.0018$) poorer in children (25.75 ± 1.33) than in adults (27.1 ± 1.21).

Effect of SNR on the comprehension and the intelligibility of the Attended stream

Figure 1 displays the comprehension scores and intelligibility ratings in the different SNR conditions in both groups.

In the Noiseless condition, all adult participants gave the maximum intelligibility rating (10), leading to a null variance. For this reason, the ANOVA for the intelligibility ratings was computed only with the 3 other conditions (+5 dB, 0 dB, and −5 dB). Doing otherwise would have violated the homoscedasticity assumption of the ANOVA.

The ANOVA performed on intelligibility rating revealed a statistically significant effect of SNR ($F_{2,76} = 119; p < 0.0001$), and a significant interaction between SNR and age-group ($F_{2,76} = 4.94; p = 0.0096$), but no significant effect of age group ($F_{1,38} = 0.05; p = 0.83$). The Lilliefors test showed that the distribution of the residuals did not deviate significantly from a normal distribution ($p = 0.15$). Post hoc comparisons performed with pairwise $t$-tests between adjacent conditions demonstrated that intelligibility rating decreased statistically significantly
from Noiseless to +5 dB (adults, \( t_{19} = 6.28, p < 0.0001 \); children, \( t_{19} = 5.06, p < 0.0001 \)), from +5 dB to 0 dB (adults, \( t_{19} = 7.19, p < 0.0001 \); children, \( t_{19} = 2.11, p = 0.048 \)) and from 0 dB to –5 dB (adults, \( t_{19} = 9.48, p < 0.0001 \); children, \( t_{19} = 5.09, p < 0.0001 \)). Comparison between adults and children revealed that children gave lower intelligibility ratings than adults in Noiseless \( (t_{38} = 2.26, p = 0.030) \) and +5 dB conditions \( (t_{38} = 2.10, p = 0.042) \) but not in the two other noisiest conditions \( (p > 0.05) \).

The ANOVA performed on comprehension scores—converted to percentage correct—revealed a significant effect of SNR \( (F_{3,114} = 27.6; p < 0.0001) \), a significant effect of age group \( (F_{1,38} = 19.4; p < 0.0001) \), and no significant interaction \( (F_{3,114} = 0.66; p = 0.58) \). The Lilliefors test showed that the distribution of the residuals did not deviate significantly from a normal distribution \( (p = 0.054) \). Comprehension scores were higher in adults \( (80.5 \pm 15.2\% \); mean ± SD across conditions and participants) than in children \( (68.6 +/- 18.7\% \), and decreased statistically significantly from Noiseless to +5 dB \( (t_{39} = 3.13, p = 0.0033) \), from +5 dB to 0 dB \( (t_{39} = 2.13, p = 0.040) \) and from 0 dB to –5 dB \( (t_{39} = 4.00, p = 0.0003) \).

---Place Figure 1 around here ---

Cortical tracking of speech in the Noiseless condition

Figures 2–4 display group-averaged coherence spectra (Fig. 2), sensor distribution (Fig. 3) and source distribution (Fig. 4) in all conditions and in both groups.

Table 1 provides the number of children and adults showing significant sensor-space \( \text{Coh}_{\text{global}}, \text{Coh}_{\text{att}}, \text{and Coh}_{\text{bckgr}} \) at <1 Hz, 1–4 Hz and 4–8 Hz frequencies.

In adults, statistically significant \( \text{Coh}_{\text{att}} \) was observed at < 1 Hz in 20/20 adults, at 1–4 Hz in 17/20 adults, and at 4–8 Hz in 17/20 adults in temporal-lobe MEG sensors in the Noiseless condition (Figs. 2 and 3). In source space, group-level coherence at <1 Hz peaked at
bilateral superior temporal sulcus (STS; left-hemisphere MNI coordinates \([-66 \ -16 \ -1]\), \(p < 0.0001\); right \([66 \ -26 \ 5]\), \(p < 0.0001\)), left inferior frontal gyrus (IFG; \([-61 \ -6 \ 34]\), \(p = 0.0037\)) and right central sulcus ([59 \ -5 \ 41], \(p = 0.034\)). In addition, \(Coh_{att}\) peaked in bilateral supratemporal auditory cortices (AC) at 1–4 Hz (left \([-65 \ -16 \ 7]\), \(p < 0.0001\); right \([64 \ -5 \ 6]\), \(p < 0.0001\)) and 4–8 Hz (left \([-65 \ -15 \ 9]\), \(p < 0.0001\); right \([64 \ -8 \ 5]\), \(p < 0.0001\)) (Fig. 4).

In children, statistically significant \(Coh_{att}\) was observed at < 1 Hz in 20/20 children, at 1–4 Hz in 10/20 children, and at 4–8 Hz in 11/20 children in temporal-lobe MEG sensors in the Noiseless condition (Figs. 2 and 3). In source space, group-level coherence at <1 Hz peaked at bilateral STS (left \([-63 \ -10 \ 6]\), \(p < 0.0001\); right \([63 \ -20 \ -3]\), \(p < 0.0001\)). It peaked at bilateral supratemporal AC at 1–4 Hz (left \([-64 \ -11 \ 13]\), \(p = 0.0016\); right \([62 \ -28 \ 4]\), \(p < 0.0001\)), and in left supratemporal AC at 4–8 Hz (\([-65 \ -17 \ 10]\), \(p = 0.018\)). Of note, coherence did peak in the right supratemporal AC at 4–8 Hz but this local maximum did not reach statistical significance ([64 \ -14 \ 5], \(p = 0.11\)).

The above results suggest that speech tracking differed somewhat in adults and children at 1–4 Hz and 4–8 Hz but not at < 1 Hz. Indeed, a smaller proportion of children than of adults showed significant tracking at 1–4 Hz (10/20 vs. 17/20; \(p = 0.041\); Fisher exact test). A similar but statistically nonsignificant trend was observed at 4–8 Hz (11/20 vs. 17/20; \(p = 0.082\)). Also, contrast between adults and children did reveal stronger tracking in adults than in children in bilateral supratemporal AC at 1–4 Hz (left \([-65 \ -16 \ 7]\), \(p = 0.0013\); right \([64 \ -5 \ 6]\), \(p = 0.0041\)) and 4–8 Hz (left \([-65 \ -15 \ 9]\), \(p < 0.0001\); right \([64 \ -8 \ 5]\), \(p < 0.0001\)), but not at < 1 Hz.

Correlation between \(Coh_{att}\) values and children’ age was statistically significant in the right STS at <1 Hz (\(r = 0.47\), \(p = 0.039\)), and in the right supratemporal AC at 4–8 Hz (\(r = 0.32\), \(p = 0.005\)).
Cortical tracking of speech in SiN conditions

In adults, group-level $Coh_{att}$ and $Coh_{global}$ maps at $< 1$ Hz displayed statistically significant ($ps < 0.05$) local maxima at bilateral STS at every SNR, in the left inferior frontal gyrus (IFG) at $0$ dB and in the left temporal pole at $-5$ dB. $Coh_{bckgr}$ was not statistically significant in any condition. At 1–4 Hz and 4–8 Hz, $Coh_{att}$ and $Coh_{global}$ maps displayed statistically significant ($ps < 0.05$) local maxima in AC bilaterally in every condition (Fig. 4). $Coh_{bckgr}$ was statistically significant at $0$ dB and $-5$ dB in right AC.

In children, group-level $Coh_{att}$ and $Coh_{global}$ maps at $< 1$ Hz displayed statistically significant ($ps < 0.05$) local maxima in bilateral STS at every SNR except at $-5$ dB where only $Coh_{att}$ displayed significant local maxima (Fig. 4). At 1–4 Hz, $Coh_{att}$ and $Coh_{global}$ maps displayed statistically significant ($ps < 0.05$) local maxima in AC bilaterally at $+5$ and $0$ dB, and only $Coh_{att}$ displayed significance coherence in right AC at $-5$ dB (Fig. 4). At 4–8 Hz, $Coh_{att}$ and $Coh_{global}$ maps displayed statistically significant ($ps < 0.05$) local maxima only at right AC at $+5$ and $0$ dB, but not at $-5$ dB (Fig. 4). $Coh_{bckgr}$ was not statistically significant in any condition and frequency band of interest.
Cortical processing of the auditory scene in SiN conditions

In adults, Coh_\text{att} was stronger than Coh_\text{global}, i.e., MEG signals tracked more the attended stream than the global scene. At < 1 Hz, cortical areas showing this effect included bilateral STS at every SNR (+5 dB, 0 dB and –5 dB; \( p < 0.0001 \)) and left inferior frontal gyrus at 0 dB. Significantly stronger Coh_\text{att} than Coh_\text{global} was found at bilateral AC at every SNR at 1–4 Hz (+5 dB, \( p = 0.0004 \); 0 dB, \( p = 0.0003 \); –5 dB, \( p = 0.0001 \)) and 4–8 Hz (+5 dB, \( p < 0.0001 \); 0 dB, \( p < 0.0001 \); –5 dB, \( p = 0.0005 \)).

In children, Coh_\text{att} at < 1 Hz was statistically significantly stronger than Coh_\text{global} in bilateral STS at every SNR (+5 dB, 0 dB and –5 dB; \( ps < 0.0001 \)) and in central opercular cortex at –5 dB. At 1–4 Hz, significantly stronger Coh_\text{att} than Coh_\text{global} was found in right STS (+5 dB, \( p = 0.0011 \); 0 dB, \( p = 0.0011 \); –5 dB, \( p = 0.0042 \)). At 4–8 Hz, significantly stronger Coh_\text{att} than Coh_\text{global} was found at –5 dB and in a non-auditory area (left superior frontal gyrus; \( p = 0.034 \)).

Comparison of the tracking in adults vs. children

The comparison of coherence values between adults and children revealed that tracking at frequencies < 1 Hz was stronger in adults than children in left inferior frontal gyrus at 0 dB (\( p = 0.030 \)) and in bilateral STS at –5 dB (left, \( p = 0.0003 \); right, \( p = 0.0056 \)). At 1–4 Hz, Coh_\text{att} was significantly higher in adults than in children in bilateral AC at every SNR (\( ps < 0.05 \)) except at 0 dB in the right AC where this effect was only marginally significant (\( p = 0.054 \)). At 4–8 Hz, Coh_\text{att} was significantly higher in adults than in children in bilateral AC at every SNR (\( ps < 0.0001 \)) (Fig. 6).

In contrast, no brain area displayed significantly stronger Coh_\text{att} in children than in adults at any frequency band of interest.
Effect of age group, SNR, and hemispheric lateralization on cortical tracking of speech

The effect of age group, SNR, and hemispheric lateralization on cortical tracking of speech was sought for with 3-way repeated measures ANOVA for the three frequency bands of interest separately. Since both adults and children’s brains track preferentially the Attended stream rather than the Global scene, the ANOVA was performed on Coh_{att} values only.

At < 1 Hz, the ANOVA revealed a significant main effect of noise level on Coh_{att} (F_{3,114} = 66.64, p < 0.0001), a significant interaction between SNR condition and age group (F_{3,114} = 4.23, p = 0.0071), and a significant interaction between SNR condition and hemispheric lateralization (F_{3,114} = 9.35, p < 0.0001), but no other significant effects (ps > 0.05). In particular, there was no significant interaction between age group and hemispheric lateralization (F_{1,38} = 0.07, p = 0.79), showing that the effect of noise on hemispheric lateralization was similar in both age groups.

Figure 7 illustrates these effects identified on cortical tracking of Attended speech stream. The main effect of noise was explained by a decrease in Coh_{att} as SNR increased. The interaction between SNR condition and age group was explained by a faster decrease in children’s than adults’ Coh_{att} as a function of decreasing SNR (see Figs. 4 and 7). Supporting this interpretation, adults had higher Coh_{att} than children at −5 dB (t_{38} = 3.88, p = 0.0004; t-test) but not at 0, +5 dB and Noiseless (ps > 0.17). The interaction between SNR condition and hemispheric lateralization was explained by a faster decrease in Coh_{att} in the right than the left STS as a function of decreasing SNR (see Figs. 4 and 6). Supporting this interpretation, post-hoc comparisons revealed that Coh_{att} at right STS decreased as soon as the
Multitalker background was added (Noiseless vs. +5 dB: $t_{19} = 4.9, p < 0.0001$) and it further diminished as SNR decreased (+5 dB vs. 0 dB, $t_{19} = 4.1, p = 0.0002$; 0 dB vs. −5 dB, $t_{19} = 4.23, p = 0.0001$). In contrast, Cohatt in the left STS decreased significantly only in the two noisiest conditions (Noiseless vs. +5 dB, $t_{19} = 0.49, p = 0.62$; +5 dB vs. 0 dB, $t_{19} = 3.11, p = 0.0035$; 0 dB vs. −5 dB, $t_{19} = 5.3, p < 0.0001$).

At 1–4 Hz, the ANOVA revealed a significant main effect of age group on Cohatt ($F_{3,114} = 10.04, p = 0.003$), a significant main effect of hemispheric lateralization ($F_{3,114} = 10.04, p = 0.003$), no significant main effect of the SNR condition ($F_{3,114} = 1.53, p = 0.21$), and no significant interactions ($ps > 0.05$). The main effect of age group was due to higher Cohatt in adults than in children. The main effect of hemispheric lateralization was explained by higher Cohatt values in the right than the left AC.

At 4–8 Hz, the ANOVA revealed a significant main effect of age group on Cohatt ($F_{3,114} = 54.64, p < 0.0001$), a significant main effect of noise level ($F_{3,114} = 13.96, p < 0.0001$), a significant effect of hemispheric lateralization ($F_{3,114} = 7.37, p = 0.0099$), a significant interaction between SNR condition and age group ($F_{3,114} = 10.08, p < 0.0001$), and no other significant interactions ($ps > 0.05$). The main effect of age group was explained by higher Cohatt in adults than in children at all SNRs ($ps < 0.001$). The interaction between SNR condition and age group was explained by the presence of SNR-related modulation in adults’ Cohatt and the absence of such modulation in children’s Cohatt (see Fig. 7). Indeed, adults’ Cohatt (mean across hemispheres) was significantly higher at intermediate SNRs (5 dB and 0 dB) than in Noiseless (5 dB, $t_{19} = 4.29, p = 0.0004$; 0 dB, $t_{19} = 2.85, p = 0.010$) and at −5 dB (5 dB, $t_{19} = 5.26, p < 0.0001$; 0 dB, $t_{19} = 5.20, p < 0.0001$), significantly higher at 5 dB than at 0 dB ($t_{19} = 2.30, p = 0.033$), and marginally higher in Noiseless than at −5 dB ($t_{19} = 1.96, p = 0.065$). The main effect of hemispheric lateralization was explained by higher or marginally
higher $Coh_{\text{att}}$ in right than left AC in every SNR condition (noiseless, $t_{39} = 2.06, p = 0.046$; $5 \text{dB}$, $t_{39} = 2.37, p = 0.023$; $0 \text{dB}$, $t_{39} = 1.93, p = 0.061$; $-5 \text{dB}$, $t_{39} = 1.71, p = 0.095$).

Discussion

This study discloses commonalities between children and adult’s cortical tracking of SiN. First, both children and adults’ auditory systems similarly tracked the attended speaker’s voice more than the global auditory scene at $< 1 \text{ Hz}$ and $1–4 \text{ Hz}$. Second, cortical tracking of the attended stream in SiN conditions was at $< 1 \text{ Hz}$ similarly left-hemisphere-dominant in children and adults. Furthermore, in both groups, the STS was the main brain area underpinning this tracking at $< 1 \text{ Hz}$. Compared with adults, children displayed (i) reduced cortical tracking of speech at $1–4 \text{ Hz}$, and particularly at $4–8 \text{ Hz}$ (even in noiseless conditions); (ii) increasing multitalker background level compromised children more than adults in cortical tracking of speech at $< 1 \text{ Hz}$, and (iii) children did not exhibit selective cortical representation of SiN at $4–8 \text{ Hz}$.

Inaccurate cortical tracking of speech at $4–8 \text{ Hz}$ in children

Studies previously reported that ongoing cortical oscillations track speech regularities, especially at syllable timescale, which corresponds to $4–8 \text{ Hz}$ frequencies (Ding and Simon, 2012b; Gross et al., 2013; Koskinen and Seppä, 2014; Ding et al., 2016). Since the strength of this $4–8 \text{ Hz}$ cortical tracking is related to speech intelligibility (Luo and Poeppel, 2007; Peelle et al., 2013; Doelling et al., 2014), it has been hypothesized to reflect active speech perception mechanisms, very likely involved in parsing incoming connected speech into discrete syllabic units (Giraud and Poeppel, 2012; Teng et al., 2017).
We demonstrated that children’s auditory system is less proficient than adults’ auditory system at tracking the attended speech envelope at 1–4 Hz and 4–8 Hz. Still, without noise, about half of children showed significant cortical tracking of the attended speech envelope at 1–4 Hz and 4–8 Hz. Crucially, cortical tracking of speech envelope at syllable rate correlated in the right AC positively with the children’s age. Furthermore, 4–8 Hz tracking was sensitive to noise intensity so that the number of children with significant tracking decreased with increasing noise level from 11 subjects in Noiseless to only 2 at −5 dB. Finally, at 4–8 Hz, children did not exhibit the selective tracking of speech in auditory areas seen in adults.

These results are at odds with at least one of the views claiming that 1) the sensitivity to syllables is acquired early in age and that 2) tracking at 4–8 Hz relates to processing of syllabic units. Several studies highlighted the early-developed ability of infants and children to discriminate syllables. Neurophysiological evidence of syllable discrimination has been found already in preterm infants (Mahmoudzadeh et al., 2013) and 8-month-old infants are able to segregate newly learned words from syllable strings (Saffran et al., 1996). Furthermore, speech slow amplitude modulations are one of the first cues that children identify when they listen to speech (Nittrouer, 2006), allowing them to detect syllable boundaries, which is mandatory for an accurate perception of speech (Goswami, 2011; Poelmans et al., 2011). However, recent findings indicate that children under 10 years are less accurate than adults at identifying syllable boundaries when these are only defined by amplitude modulations in temporal envelope (Cameron et al., 2018), and that theta band cortical tracking is not speech-specific (Molinaro and Lizarazu, 2017). Also questioning the link between syllable processing and the 4–8-Hz tracking is the consistent finding that such coupling is right-hemisphere dominant (Luo and Poeppel, 2007; Giraud and Poeppel, 2012; Gross et al., 2013; Peelle et al., 2013). In this context, our results argue for progressive
evolution from childhood to adulthood of abilities to track the acoustic envelope of speech at 4–8 Hz.

Noise easily corrupts cortical tracking of speech in children

In adverse listening conditions (i.e., −5 dB), auditory system lost, in a substantial proportion of children (70%), the capability to track the Attended stream at < 1 Hz whereas 80% of adults exhibited significant tracking in this condition. Since children’s SiN behavioural performances are typically lower than those of adults (Elliott, 1979), we can postulate that the poor performance was at least partially related to a limited central auditory capacity to segregate the Attended stream from the Multitalker background at < 1 Hz when SNR decreased. This ability of the auditory system likely improves during adolescence, given the outcome in young adults reported here and elsewhere (Ding and Simon, 2012a, 2013b; Vander Ghinst et al., 2016). Hence, our study argues for a developmental origin of the selective cortical tracking of the Attended stream at < 1 Hz. These data are in line with our results and previous psychoacoustic studies that demonstrated children’s poor speech comprehension in adverse listening conditions (Johnson, 2000; Talarico et al., 2007; Neuman et al., 2010; Klatte et al., 2013).

Children’s and adults’ auditory system tracks the attended speech stream in SiN conditions at < 1 Hz and at 1–4 Hz.

At < 1 Hz and 1–4 Hz, coupling between envelopes of the listened sounds and the activity of non-primary auditory cortex was stronger with the Attended stream than with the Global scene, both in children and adults. This finding is in line with former studies performed in adults (Ding and Simon, 2012a; Mesgarani and Chang, 2012; Ding and Simon, 2013b; Zion Golumbic et al., 2013; Vander Ghinst et al., 2016). Yet, and in contradiction with
our initial hypothesis, the current study is the first to demonstrate that this selective tracking already exists in children, at least up to a certain noise level. The preferential tracking of the slowest (< 1 Hz) speech modulations took place, like in adults, at bilateral STS, demonstrating that this brain area extracts or has a preferential access to the Attended stream in reasonable SNR conditions. Still, the preferential tracking of the Attended stream at left IFG at 0 dB was higher in adults than in children. Moreover, the absence of specific tracking of the Multitalker background argues for an object-based neural coding of the attended speaker’s voice in children’s higher-order auditory cortical areas up to a certain SNR level (Simon, 2014; Puvvada and Simon, 2017). Interestingly, at 1–4 Hz, this selective tracking occurred in bilateral AC in adults but only in right AC in children. Since recent findings have shown that perceptually relevant word segmentation takes place in left temporal cortex (Keitel et al., 2018), the lack of selective cortical tracking of speech at word repetition rate (1–4 Hz) in left temporal cortex could partly explain SiN difficulties in children.

Effects of the multitalker background on hemispheric lateralization of cortical tracking of speech at < 1 Hz

As demonstrated here and elsewhere (Power et al., 2012; Vander Ghinst et al., 2016; Destoky et al., 2019), the left-hemisphere cortical tracking of speech at < 1 Hz is essentially preserved in a multitalker background up to a SNR of 0 dB, while it is compromised in the right hemisphere as soon as a background noise is added. Ours is the first study to demonstrate that this hemispheric asymmetry in cortical tracking occurred similarly in children and in adults, but with the noticeable difference that children lost the tracking in the more challenging conditions.

Left-hemisphere-dominant noise-resistant cortical tracking of speech at STS (and IFG in adults) could imply an active cognitive process that promotes speech recognition
(Schroeder et al., 2008; Schroeder and Lakatos, 2009; Power et al., 2012) through increased access to lexical and semantic representations (Binder et al., 2009; Liebenthal et al., 2014). This left-lateralized process is likely related to correct identification and comprehension of the targeted auditory stream (Alain et al., 2005; Bishop and Miller, 2009). Since the coupling between cortical oscillations and the low-frequency rhythmic structure of an attended acoustic stream seems to be under attentional control (Lakatos et al., 2013), the differential hemispheric effect of noise on cortical tracking of speech could be related to mechanisms of selective attention. As noise impairs children more strongly than adults in addition to auditory- and speech-related tasks also in non-auditory cognitive processes, such as reading and writing (for a review, see Klatte et al., 2013), we can hypothesize that the detrimental effect of acute and chronic noise exposure on different cognitive functions is at least partially related to the crucial attentional load needed to understand speech in adverse hearing environment.

Conclusion

The ability of children’s brains to track speech temporal envelopes at syllable rate (4–8 Hz) was drastically reduced in comparison with adults regardless of the SNR. Similarly as adults, children displayed stronger tracking of the Attended stream than of the Global scene in SiN conditions at phrasal and word rates, but their tracking ability was more easily corrupted by increasing noise. Children’s poor SiN comprehension performances were therefore likely related to a limited central auditory capacity to segregate the Attended stream from the Multitalker background at phrasal and word rates as a function of decreasing SNR and at 4–8 Hz regardless of the SNR. These results further elucidate the neurophysiological mechanisms accounting for children’s difficulties to adequately segregate speech in informational masking conditions.
References

thalamo-cortical network implicated in successful speech separation and identification.
NeuroImage 26:592-599.

Brain Mapp 7:254-266.


Berman S, Friedman D (1995) The development of selective attention as reflected by event-

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A
critical review and meta-analysis of 120 functional neuroimaging studies. Cereb
Cortex 19: 2767-2796.

Bishop CW, Miller LM (2009) A multisensory cortical network for understanding speech in

coherence estimated with segment overlapping. Signal Processing 87:1100-1117.

coherence mainly reflects movement-induced proprioceptive feedback. NeuroImage
106:382-390.

(2012) Neuronal network coherent with hand kinematics during fast repetitive hand


temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. J Neurosci 36:1596-1606.


Figure 1: Comprehension scores (left graph) and intelligibility ratings (right graph) in adults (black) and children (grey). Bars indicate the mean and range. Comprehension scores are reported in percentage of questions (16 for adults and 8 for children) answered correctly, and intelligibility ratings ranged from 0 (totally unintelligible) to 10 (perfectly intelligible). The comprehension and intelligibility of the Attended stream decreased significantly as SNR decreased. Horizontal brackets indicate the outcome of post-hoc paired t-tests between adjacent conditions (*** p < 0.001, ** p < 0.01, * p < 0.05).
Figure 2: Spectra of cortical tracking of speech in the 4 speech-in-noise conditions and corresponding sound excerpts. Group-averaged coherence spectra are shown separately for adults (left column) and children (right column), and when estimated with the temporal envelope of the different components of the auditory scene: the Attended stream (black connected trace), the Multitalker background (gray connected trace), and the Global scene (gray dotted trace). Each spectrum represents the mean across subjects (12 in each age group) of the maximum coherence across all sensors. The sound excerpts showcase the Attended stream (black traces) and the Multitalker background (gray traces) and their relative amplitude depending on the signal to noise ratio.
Figure 3: Cortical tracking of speech in the sensor-space. Plots display the group-averaged coherence distributions obtained based on gradiometer data, implying that maximum coherence should peak right above generating brain sources. There is one distribution map for each frequency band of interest (< 1 Hz, 1–4 Hz, 4–8 Hz), age group (adults, children) component of the auditory scene (Attended stream, Global scene), and SNR condition (Noiseless, +5 dB, 0 dB, –5 dB).
Figure 4: Cortical tracking of the attended speech stream at < 1 Hz, 1–4 Hz, and 4–8 Hz. The group-level coherence maps (n = 20 in each map) were masked statistically above the significance level (non-parametric permutation statistics). One source distribution is displayed for each possible combination of age group (adults, top panel; children, bottom panel) and SNR condition (from left to right, Noiseless, +5 dB, 0 dB, and −5 dB).
Figure 5: Correlation between children’s age and their speech tracking values, depicted by the peak group-level coherence, in the noiseless condition at the right superior temporal sulcus at < 1 Hz and supratemporal auditory cortex at 4–8 Hz.
Figure 6: Contrasts between cortical tracking of the attended speech ($Coh_{att}$) in adults versus children ($Coh_{att, adults} - Coh_{att, children}$) at < 1 Hz, 1–4 Hz, and 4–8 Hz in all SNR conditions (Noiseless, +5 dB, 0 dB, and –5 dB). The group-level difference coherence maps ($n = 20$ in each map) were masked statistically above the significance level.
Figure 7: Illustration of the interaction effects identified on speech brain tracking of the attend speech stream in adults (left column) and children (right column), at <1 Hz, 1–4 Hz and 4–8 Hz. Bars indicate the mean and standard error on the mean of $f$-transformed (with $f(\cdot) = \tanh(\sqrt{\cdot})$) coherence values at the right (solid lines) and left (dashed lines) cortical area of peak group-level coherence (superior temporal sulcus at <1 Hz, and supratemporal auditory cortex at 1–4 Hz and 4–8 Hz).
Table 1: Number of adults and children showing statistically significant coherence (surrogate-data-based statistics) in at least one sensor for each audio signal, condition, and frequency band of interest.
Attended stream | Multitalker background | Global scene
The graphs show the coherence in two frequency bands: 1 Hz and 4-8 Hz, as a function of age (in months). For the 1 Hz band, the correlation coefficient is $r = 0.47$. For the 4-8 Hz band, the correlation coefficient is $r = 0.50$. The data points are plotted on a linear scale, and the trend lines are also linear.
Noiseless +5 dB 0 dB –5 dB

arctanh√(coherence)

CHILDREN

ADULTS

< 1 Hz

1–4 Hz

4–8 Hz
<table>
<thead>
<tr>
<th>Condition</th>
<th>Attended stream</th>
<th>Multitalker background</th>
<th>Global scene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adult</td>
<td>Children</td>
<td>Adult</td>
</tr>
<tr>
<td>&lt;1 Hz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noiseless</td>
<td>20</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>+5 dB</td>
<td>20</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>0 dB</td>
<td>20</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>−5 dB</td>
<td>16</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>1–4 Hz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noiseless</td>
<td>17</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>+5 dB</td>
<td>16</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>0 dB</td>
<td>16</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>−5 dB</td>
<td>10</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>4–8 Hz</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noiseless</td>
<td>17</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>+5 dB</td>
<td>20</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>0 dB</td>
<td>19</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>−5 dB</td>
<td>15</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>