Article

Current Biology

Voice categorization in the four-month-old human brain

Highlights

- Selective brain responses to voices emerge as early as 4 months of age
- Voice selectivity in infants is partially independent of acoustic features
- The topography of voice selectivity in infants resembles the adult one
- The infant brain is endowed with selectivity for socially relevant sounds

Authors

Roberta P. Calce, Diane Rekow, Francesca M. Barbero, Anna Kiseleva, Siddharth Talwar, Arnaud Leleu, Olivier Collignon

Correspondence

roberta.calce95@gmail.com (R.P.C.), olivier.collignon@uclouvain.be (O.C.)

In brief

Calce et al. rely on a highly sensitive paradigm combining fast periodic auditory stimulation (FPAS) with scalp electroencephalography (EEG) to demonstrate that the infant brain produces a reliable preferential response to voices early in life.

Calce et al., 2024, Current Biology *34*, 46–55
January 8, 2024
https://doi.org/10.1016/j.cub.2023.11.042

CellPress

CellPress

Current Biology

Article

Voice categorization in the four-month-old human brain

Roberta P. Calce,[1,5,*] Diane Rekow,[2,3] Francesca M. Barbero,[1] Anna Kiseleva,[2] Siddharth Talwar,[1] Arnaud Leleu,[2] and Olivier Collignon[1,4,6,*]

[1]Crossmodal Perception and Plasticity Laboratory, Institute of Research in Psychology (IPSY) and Institute of Neuroscience (IoNS), Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

[2]Development of Olfactory Communication and Cognition Lab, Centre des Sciences du Goût et de l'Alimentation, Université Bourgogne Franche-Comté, Université de Bourgogne, CNRS, Inrae, Institut Agro Dijon, 21000 Dijon, France

[3]Biological Psychology and Neuropsychology, University of Hamburg, 20146 Hamburg, Germany

[4]School of Health Sciences, HES-SO Valais-Wallis, The Sense Innovation and Research Center, 1007 Lausanne & Sion, Switzerland

[5]X (formerly Twitter): @roberta_calce

[6]Lead contact

*Correspondence: roberta.calce95@gmail.com (R.P.C.), olivier.collignon@uclouvain.be (O.C.)

https://doi.org/10.1016/j.cub.2023.11.042

SUMMARY

Voices are the most relevant social sounds for humans and therefore have crucial adaptive value in development. Neuroimaging studies in adults have demonstrated the existence of regions in the superior temporal sulcus that respond preferentially to voices. Yet, whether voices represent a functionally specific category in the young infant's mind is largely unknown. We developed a highly sensitive paradigm relying on fast periodic auditory stimulation (FPAS) combined with scalp electroencephalography (EEG) to demonstrate that the infant brain implements a reliable preferential response to voices early in life. Twenty-three 4-month-old infants listened to sequences containing non-vocal sounds from different categories presented at 3.33 Hz, with highly heterogeneous vocal sounds appearing every third stimulus (1.11 Hz). We were able to isolate a voice-selective response over temporal regions, and individual voice-selective responses were found in most infants within only a few minutes of stimulation. This selective response was significantly reduced for the same frequency-scrambled sounds, indicating that voice selectivity is not simply driven by the envelope and the spectral content of the sounds. Such a robust selective response to voices as early as 4 months of age suggests that the infant brain is endowed with the ability to rapidly develop a functional selectivity to this socially relevant category of sounds.

INTRODUCTION

Voices are arguably the most relevant sound for social interactions in humans, providing listeners not only with linguistic information but also with a wealth of para-linguistic information about speakers, such as their identity, sex, age, emotional status, and trustworthiness.[1–3] A fundamental perceptual ability in humans is to successfully segregate voices from the incoming stream of auditory information and to jointly classify as voices highly heterogeneous voice exemplars heard under variable listening conditions.[4,5] Despite this perceptual challenge, human voices are seamlessly and rapidly categorized among many other sounds by adults.[6–8] Efficient voice categorization is supported by voice-selective brain regions located along the bilateral superior temporal sulcus as well as the superior temporal gyrus, referred to as the temporal voice areas (TVAs),[1,9] and whose sensitivity to voice stimuli is not fully accounted for by acoustic features[9,10] or speech content.[11] Interestingly, cortical regions responding preferentially to species-specific vocalizations have also been found in macaque monkeys[12] and dogs,[13] suggesting a phylogenetic origin of the neurobiology of voice selectivity in humans.[14]

However, the ontogeny of voice categorization in the human brain remains largely unknown. Behavioral studies have shown that even before birth, fetuses are able to differentiate between familiar and unfamiliar voices,[15,16] and electroencephalography (EEG) studies have demonstrated that, compared with strangers' voices, young infants respond preferentially to a familiar voice (e.g., their mother's[17–19] or their father's voice[20]). However, these studies do not speak about perceptual or brain selectivity to vocal sounds per se[21] because familiarity detection is not specific to voices but can be detected in infants with any type of sound.[22] In addition, when investigating voice processing in young infants, it is often the case that stimuli relied on very specific voice exemplars, such as lullabies[23] or recordings in "motherese,"[24,25] leaving unanswered whether a categorical response to natural voices that is partially independent from simple acoustic features emerges early in life.[26–28] One pioneering study[29] examined the development of the preferential response to voices when compared with non-vocal sounds during infancy using near-infrared spectroscopy (fNIRS). Even though vocal stimuli were confounded by speech (including words and non-words) and had important acoustic differences with non-vocal sounds,[9] the study showed that 7-month-olds, but not 4-month-olds, had

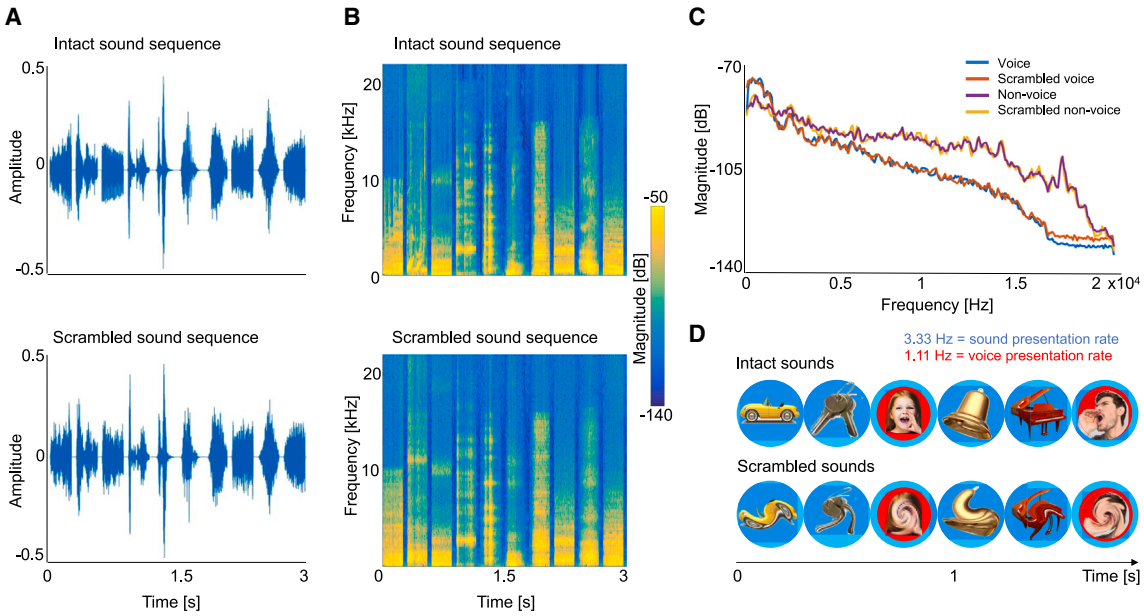

Figure 1. Experimental design

(A) 3-s excerpts of the sequences (10 sounds) for the intact (top) and scrambled (bottom) conditions. Note that the sounds are presented in the same order in both conditions for illustration but were randomly presented within each sequence during the experiment (see STAR Methods).

(B) Spectrograms of the 3-s excerpts for the intact (top) and scrambled (bottom) sequences.

(C) Bode magnitude plot of the magnitude in decibels for the averaged sounds of vocal and non-vocal stimuli for intact and scrambled sequences.

(D) Schematic representation of auditory stimulation in intact (top) and scrambled (bottom) conditions. 250-ms sounds are embedded together with a 50-ms silence gap between each sound. Sequences are created by presenting general sounds at 3.33 Hz (blue circles) with voices systematically presented at 1.11 Hz (red circles).

increased responses in the superior temporal cortex to the human voice.[29] This indication of a relatively late development of a preferential brain response to voices contrasts with the observation that 4-month-old infants already have a preferential response to natural and complex face images.[30] This is particularly intriguing given that the auditory system develops earlier than the visual one[31] and that infants similarly attend to faces and voices very early in life.[32]

The absence of systematic control conditions (i.e., contrasting unfamiliar voices to other sound categories and to acoustically controlled sounds) and the challenge of interpreting absence of evidence in young infants[19] are linked to the well-known methodological challenges of developmental research.[21,33] Neuroimaging studies are limited in designing optimal infant-friendly studies, with the consequence that often a large amount of data is excluded[34] or that the testing is often limited to asleep infants.[26] The progress of event-related potential recordings with EEG paradigms has largely increased the feasibility of developmental studies[35]; yet, the number of participants excluded and the time needed to collect reliable data across a variety of control conditions are still a current challenge.[36] Developmental research therefore calls for a paradigm independent from behavioral performance and that can generate data with high signal-to-noise within a short testing time and without contamination by irrelevant attentional, motor, or decisional processes.[37]

In the present study, we reveal that 4-month-old infants have the neural architecture supporting advanced voice categorization processes. To do so, we adapted a fast periodic auditory stimulation (FPAS) paradigm that relies on the principles of Frequency-Tagging EEG that was initially developed to tag voice selectivity in adults.[7] This approach relies on the fact that a periodic modulation of a sensory feature elicits a periodic variation in brain activity that projects at the frequency of the periodic modulation and its harmonics (i.e., integer multiples) in the EEG spectrum.[38,39] The FPAS approach can tag two key perceptual abilities to assess mature categorical representation in the brain: *discrimination* of vocal from non-vocal stimuli and *generalization* of this selective response to voices across a variety of vocal exemplars.[7] Here, twenty-three 4-month-old infants listened to a stream of various sounds presented at a rate of 3.33 Hz in which heterogeneous voices were inserted in a periodic fashion at 1.11 Hz. Infants were also presented with a scrambled version of the sounds, where the frequency content and envelope are preserved, but recognizability is disrupted[40] (Figure 1). We found that the infant brain can discriminate and generalize short voice segments included in a rapid stream of various non-vocal sounds, leading to a robust voice-selective response over posterior and temporal regions. Importantly, a significant difference was found between the neural activity elicited by intact and scrambled voices, indicating that the voice-selective response cannot be explained only by the spectrogram and the envelope of the voice excerpts and suggesting high-level voice categorization in the infant brain as early as 4 months of age.

RESULTS

Visual inspection of the grand-averaged amplitude spectrum pooled across channels for the intact condition indicates that

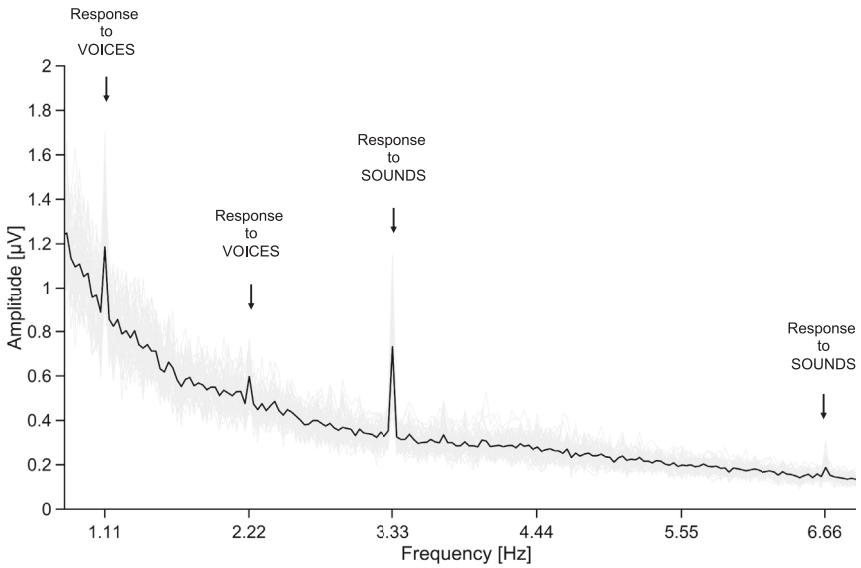

Figure 2. EEG amplitude spectrum for the intact condition

Grand-averaged raw amplitude spectrum for the intact sound sequence condition. Light gray lines depict each of the 63 channels, whereas the black line represents the average of all EEG channels. A response at the sound presentation rate (3.33 Hz) and its first harmonic (6.66 Hz) is clearly visible in the EEG spectrum. This response indicates that the brain segments the sound units that compose the sequences. The voice-selective response is visible in the EEG spectrum at the voice presentation frequency (1.11 Hz) and its first harmonic (2.22 Hz).

presenting a variety of recognizable sounds at a rapid rate of 3.33 Hz elicits a clear general response at the same frequency and its harmonics in the 4-month-old infant brain (Figure 2). Importantly, a voice-selective response is also clearly visible at the rate of voice presentation (1.11 Hz) and subsequent harmonic (2.22 Hz), revealing a differential neural activity elicited by voices compared with non-vocal sounds that generalizes across voice exemplars. After having compiled each neural response across significant harmonics (see section frequency-domain analysis in the method details), we investigated their significance over the whole scalp and whether they differ between intact and scrambled conditions.

General auditory response

The significance of the general auditory response was first determined in the intact and scrambled conditions separately using *Z* scores contrasting the amplitude of the response to the mean amplitude of background noise ($p < 0.05$, false discovery rate [FDR]-corrected for the entire electrode set, one-tailed, signal > noise). We observed a strong response distributed over the whole scalp, as most channels show a significant response in both conditions (all 63 channels for the intact sounds and 61 channels for the scrambled sounds) (Table S2). Signal-to-noise ratios (SNRs) calculated on the average of all channels reveals a 67% and 46% signal increase compared with surrounding noise in the intact (SNR = 1.67) and scrambled (SNR = 1.46) conditions, respectively (Figure 3A). Baseline-corrected amplitudes reveal a topographical distribution similar for each condition, being larger at bilateral occipito-temporal channels, with the maximum amplitude for channel P10 in both conditions (mean ± SEM: 1.18 ± 0.2 μV in intact, 0.96 ± 0.1 μV in scrambled) (Figure 3B). When the two conditions were compared by calculating *Z* scores on their amplitude difference ($p < 0.05$, FDR-corrected for the entire electrode set, two-tailed, intact ≠ scrambled), a significantly higher response was found in the intact condition than in the scrambled condition for 37 out of the 63 channels (≈59%), mainly located over occipital and temporal regions. In contrast, no channel yielded a significantly larger response in the scrambled condition (Table S2).

Voice-selective response

As for the general auditory response, the significance of the voice-selective response was first investigated in the intact and scrambled conditions separately. For the intact condition, three contiguous electrodes showed a significant response over the left temporal region (C5: $Z = 4.05$, $p_{FDR} = 0.002$; FC5: $Z = 3.37$, $p_{FDR} = 0.008$; T7: $Z = 3.2$, $p_{FDR} = 0.011$) together with another three contiguous electrodes at right occipito-temporal scalp locations (P8: $Z = 3.62$, $p_{FDR} = 0.005$; PO8: $Z = 2.99$, $p_{FDR} = 0.016$; PO6: $Z = 2.97$, $p_{FDR} = 0.016$. For the scrambled condition, only two temporal electrodes showed a significant response, one in each hemisphere and with no overlap with the channels identified in the intact condition (TP7: $Z = 3.05$, $p_{FDR} = 0.036$; TP8: $Z = 3.41$, $p_{FDR} = 0.021$; Figure 4A). For the intact condition, the magnitude of the response is maximal at channel C5 over the left temporal region (0.99 ± 0.24 μV) and at channel P8 over the right occipito-temporal region (0.88 ± 0.23 μV). For the scrambled condition, the magnitude of the voice-selective response is 0.51 ± 0.13 μV at channel TP7 and 0.62 ± 0.17 μV at channel TP8 for the left and the right areas, respectively. The mean baseline-corrected amplitude across these responding channels is 0.80 ± 0.17 and 0.56 ± 0.12 μV for the intact and scrambled conditions, respectively (Figure 4B).

Next, to compare the strength of the voice-selective response between conditions, we computed *Z* scores on their amplitude difference and found five electrodes with a significantly larger response to intact rather than scrambled voices (Figure 4A), including two electrodes over the left temporal region (FC5: $Z = 3.39$, $p_{FDR} = 0.016$; C5: $Z = 5.4$, $p_{FDR} < 0.001$) and three electrodes at right occipito-temporal sites (P8: $Z = 2.93$, $p_{FDR} = 0.049$; PO8: $Z = 2.88$, $p_{FDR} = 0.049$, O2: $Z = 3.36$, $p_{FDR} = 0.016$). In contrast, no electrode showed a significantly larger voice-selective response for the scrambled condition (Table S3). SNR values of 1.61 and 1.31 (61% and 31% of signal increase) were respectively obtained for the intact and scrambled conditions in these voice-selective channels (Figure 4C). *Z* scores were also computed on the individual amplitude difference and 19 out of 23 infants showed a significantly larger response to intact rather than scrambled voices in at least one out of the 63 channels ($p < 0.005$, two-tailed,

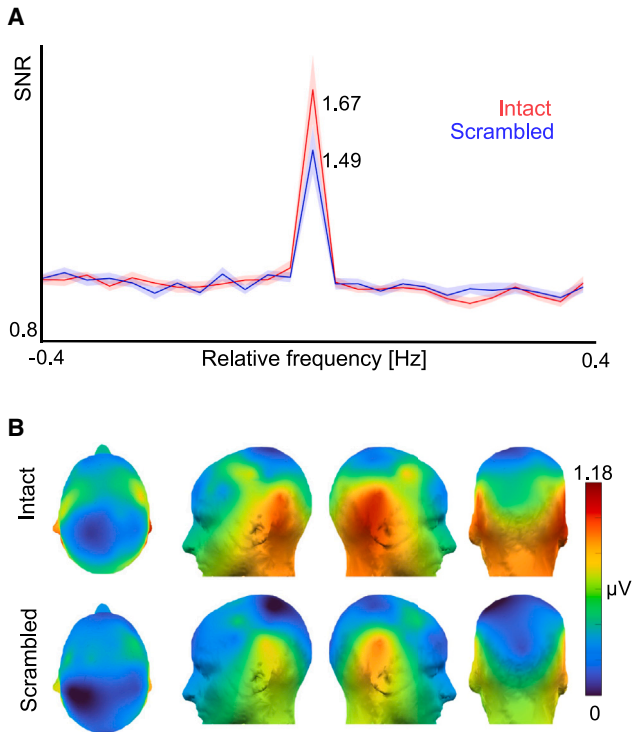

Figure 3. General auditory response for the intact and scrambled sounds

(A) Signal-to-noise ratio graph of the averaged channels in the intact (red) and scrambled (blue) conditions for the sum of the seven consecutively significant harmonics (i.e., up to 23.31 Hz). Shaded areas depict the standard error of the mean (SEM).

(B) Amplitude distribution of the summed response to intact (top) and scrambled (bottom) sound sequences. Topographic head plots were obtained by summing the baseline-corrected amplitude of the seven consecutively significant harmonics (i.e., up to 23.31 Hz).

See Table S2 for *Z* scores at each channel.

intact ≠ scrambled, Table S4). Even when applying a highly conservative threshold at the single-subject level (FDR correction for the entire electrode set), 15 out of 23 infants showed an enhanced response to intact versus scrambled voices ($p_{FDR} < 0.05$, two-tailed, intact ≠ scrambled) (Table S5).

To ensure that the observed advantage for the intact condition remains when accounting for the variability of the response across infants, we then compared the baseline-corrected amplitude of the voice-selective response between conditions, considering individual infant data. First, we considered the mean baseline-corrected amplitude across the eight channels previously identified as showing a significant response in each condition (i.e., respectively six and two for the intact and scrambled conditions, see above) and compared them using a two-tailed paired t test ($p < 0.05$, intact ≠ scrambled). We found that the intact condition yields a significantly larger response (0.72 ± 0.15 μV) than the scrambled condition (0.42 ± 0.10 μV, $t_{(22)} = 2.13$, $p = 0.022$, Cohen's $d = 0.534$). We further confirmed the analysis using only the two best channels (i.e., presenting the highest *Z* score; one per condition) for each scalp region (means across C5 and TP7 in the left hemisphere and across P8 and TP8 in the right hemisphere) to include the same number of channels per condition. The difference was again significant in the left temporal region (intact: 0.74 ± 0.15 μV, scrambled: 0.44 ± 0.11 μV, $t_{(22)} = 2.15$, $p = 0.021$, Cohen's $d = 0.486$), while non-significant in the right occipito-temporal region despite a still-stronger voice-selective response to intact voices (0.69 ± 0.21 μV) than scrambled voices (0.56 ± 0.14 μV, $t_{(22)} = 0.77$, $p = 0.45$, Cohen's $d = 0.197$). This apparent hemispheric asymmetry of the voice-selective response is, however, not supported by the mean lateralization index calculated on the significant channels in the standard condition (−12.6 ± 12%), which was not significantly different from 0 ($t_{(22)} = 0.16$, $p = 0.157$).

Visual inspection of individual infant data revealed a focal voice-selective response with a variable topographical distribution across infants (Figure 5). To determine whether the advantage for the intact condition remains when accounting for this variable scalp topography, we considered the channel presenting the largest individual baseline-corrected amplitude in each condition and for each infant. When the best channel in the intact condition is compared with the best channel in the scrambled condition (which can be different across conditions and infants), the voice-selective response remains significantly larger in the intact (2.42 ± 0.23 μV) than the scrambled condition (1.84 ± 0.11 μV, $t_{(22)} = 2.60$, $p = 0.008$, Cohen's $d = 0.715$). This finding remains unchanged when considering the mean magnitude across the two channels (same channels across conditions, which can be different across infants), with a larger response in the intact (1.58 ± 0.19 μV) than the scrambled condition (1.20 ± 0.13 μV, paired t test, $t_{(22)} = 2.43$, $p = 0.024$, Cohen's $d = 0.269$). In sum, the higher response in the intact condition is consistent across infants and analytical choices (Figure 5).

Finally, we conducted two additional analyses to reinforce the robustness of the difference between the responses to intact and scrambled voices. First, whole-scalp cluster-based permutation testing confirmed that a cluster of two left temporal channels respond more to intact than scrambled voices (FC5: $t = 3.06$, $p = 0.006$; C5: $t = 2.39$, $p = 0.026$). Then, intact and scrambled conditions were successfully decoded from the amplitude of the voice-selective response over the whole scalp with an accuracy of 57 ± 0.03%, which was significantly above chance level ($p = 0.011$).

Voice-selective response normalization by the general auditory response

Because the general auditory response is of a larger amplitude for the intact than the scrambled condition, it could relate to different levels of attention to auditory sequences according to the nature of the stimuli and influence the strength of the voice-selective response. Thus, we normalized the voice-selective response by the general auditory response to ensure that the amplitude difference between conditions for the former is not accounted for by the amplitude difference for the latter. For each channel, individual data were normalized using the raw amplitude of the voice-selective response divided by the mean scalp-wide amplitude of the general response used as norm in the intact (1.39 μV) and scrambled conditions (1.26 μV), respectively. Resulting individual amplitudes were grand-averaged by condition and *Z* scores were calculated on the amplitude difference between conditions for all channels. Channels C5 and O2, as found in the previous analysis, were again

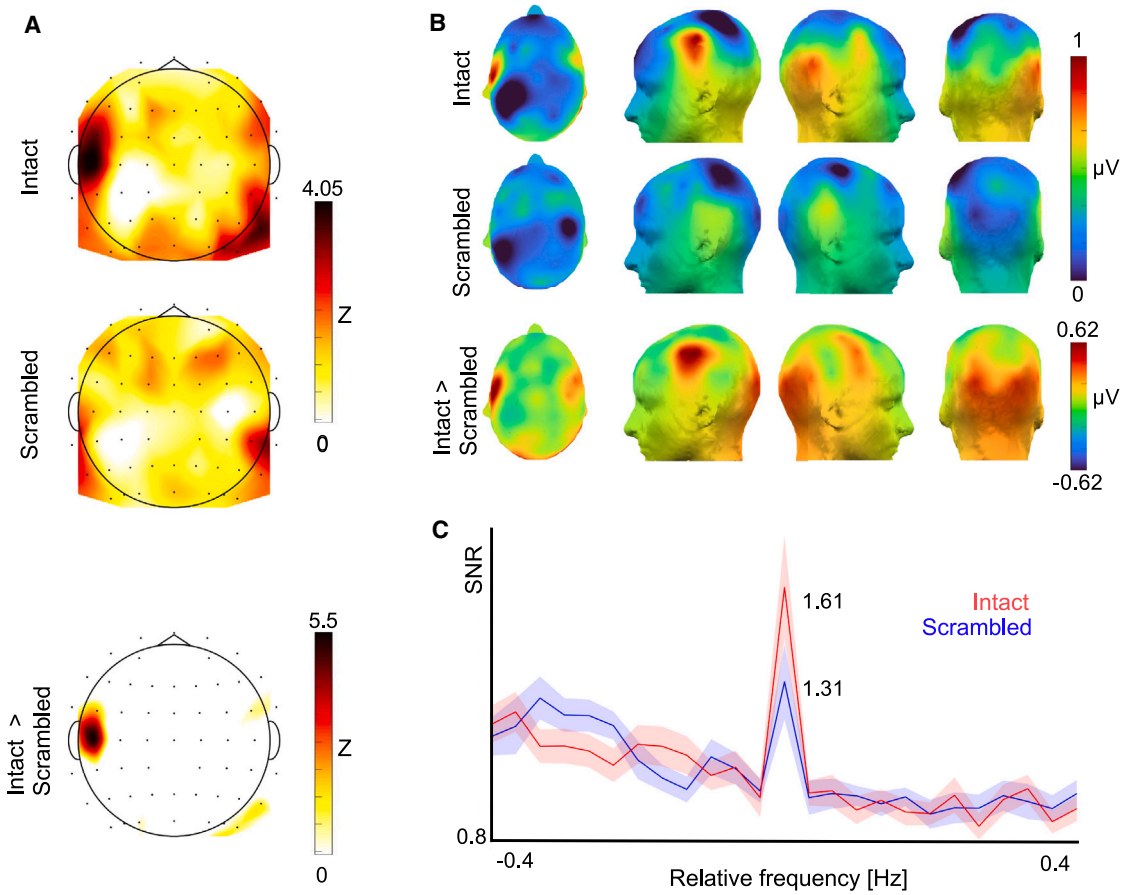

Figure 4. Voice-selective response

(A) Topographic distribution of the significant electrodes of the voice-selective response. Two-dimensional (2D) head plots represent the *Z* score values of the intact (top), scrambled (center), and intact minus scrambled (bottom) conditions, contrasted with zero. See Table S3 for *Z* scores at each channel.

(B) Amplitude distribution of the response at intact (top), scrambled (center), and intact minus scrambled conditions. Topographic head plots were obtained by summing the baseline-corrected amplitude at the two significant harmonics (i.e., until 2.22 Hz).

(C) Signal-to-noise ratio graph of the averaged significant channels in the intact (red) and the scrambled (blue) conditions. Shaded areas depict the SEM.

significant (C5: $Z = 4.81$, $p_{FDR} < 0.001$; O2: $Z = 3.44$, $p_{FDR} = 0.018$). This shows that the voice-selective response elicited by intact voices remains higher than the one elicited by their scrambled versions, even when normalized by the amplitude of the general auditory response. Moreover, when the same procedure is applied on the mean baseline-corrected amplitude of the response across the eight responding channels previously identified, the difference between conditions remains significant, with an advantage for the intact (0.55 ± 0.1 μV) over the scrambled condition (0.33 ± 0.08 μV, $t_{(22)} = 1.79$, $p = 0.043$, Cohen's $d = 0.445$).

DISCUSSION

Given the importance of vocal information in the social development of humans and other animals, evolution may have equipped the brain with the ability to efficiently categorize voices as a specific sound category very early in development. In this study, we demonstrate that 4-month-old infants have already the neural architecture to automatically discriminate voices as a specific auditory category, partially independently from low-level acoustic features characterizing voices.

To do so, we combined EEG recordings with an FPAS approach that tags the ability of the infant brain to both *discriminate* vocal from non-vocal stimuli and *generalize* this response to voices across a variety of vocal exemplars. In recent years, the frequency-tagging EEG approach has been increasingly used to study visual categorization (e.g., faces,[41] words,[42] body limbs and houses[43]) and voice categorization in the adult brain[7]. In particular, while behavioral studies are often unable to disentangle the internal capacities from superficial constraints, as cognition in preverbal human infants must be inferred from overt motor behaviors,[44,45] frequency-tagging EEG has emerged as a powerful tool to investigate visual categorization abilities in non-verbal developmental populations[30,46–48] as it does not require an overt behavioral response. In addition, it allows the isolation of clear category-selective responses with relative immunity to artifacts and high SNR in a short acquisition time.[49] For instance, while the spontaneous movement of the infants, such as sucking or other rhythmic movements, generally

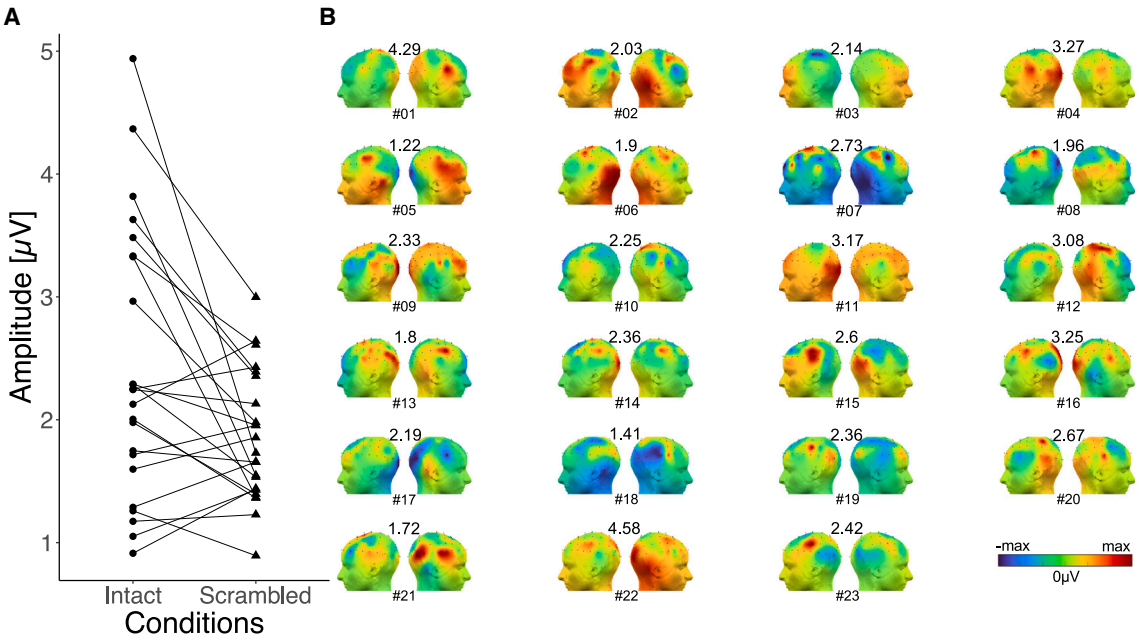

Figure 5. Individual voice-selective responses

(A) For each participant, the channels with the highest amplitude values are displayed in both conditions. The amplitude value was obtained by summing the baseline-corrected amplitude for the two harmonics of the voice frequency stimulation (1.11 and 2.22 Hz).

(B) Individual subtracted responses between the two conditions (intact minus scrambled). The maximum amplitude value of each participant (#1 to #23) used for individual scaling is indicated above its respective topography.

See Tables S4 and S5 for *Z* scores at each significant channel and participant.

elicit strong artifacts and lead to reject a lot of data in EEG infant studies, they only have a negligible influence at the frequencies of interest as they are not periodic enough to elicit a response in a narrow frequency bin. Thanks to the high frequency resolution in our study (0.0327 Hz), brain responses are captured in tiny frequency ranges while noise spreads more widely. Baseline correction from surrounding frequency bins therefore suffices to strongly reduce the contribution of background noise to the tagged responses. Finally, frequency-tagged EEG responses are defined objectively (measured at pre-experimentally defined frequencies), making the approach robust and reliable despite variable neural responses across young individuals. FPAS is thus a promising emerging tool for infant research because perceptual abilities can be inferred automatically from brain responses.[50] This technique can be used in future studies to further investigate whether categories of sounds other than the voice are present in the infant brain (e.g., for tools, musical instruments[9,51–53]) and whether those putative category-selective responses are elicited by separate neural generators.

Our results show a strong general auditory response to the rapid streams of sounds in both intact and scrambled conditions, with a high amplitude at frontal and temporal channels, and a similar number of significant channels (63 for intact, 61 for scrambled). This very robust response, with almost 100% of responsive channels, indicates that the infant brain is able to strongly synchronize to complex and heterogeneous sounds embedded together every 300 ms, in line with findings on rhythm perception using pure tones.[54,55] Studies on language processing and acquisition have mainly presented auditory stimuli in isolation,[56] whereas here we show the ability of the infant brain to segment a rapid stream of sounds embedded together as in real life, a crucial skill that needs to be acquired before language.[57] Moreover, the scalp distribution is comparable to the one found with the same approach in adults,[7] suggesting that a similar underlying architecture for processing object sounds is already present in infancy.[58] Although the responses to intact and scrambled sounds are both strongly significant, a higher activation was found in the intact condition when directly compared with the scrambled condition. This indicates that while both conditions successfully synchronized the infants' auditory system, sequences of intact and therefore recognizable sounds may have engaged more attention in the young infants.[59] Finally, the voice-selective response may have contributed to the general auditory response at 3.33 Hz, this frequency being the third harmonic of the voice-selective response. The advantage of the intact sounds at the general auditory presentation rate could therefore be partially attributed to a higher response for the intact voices over the scrambled voices.

Most importantly, we also measured voice-selective neural activity over the left temporal and right occipito-temporal electrodes. This voice-selective response is an ecologically valid measure of voice categorization under the challenging hearing conditions experienced in natural auditory environments. Indeed, as a variety of vocal and non-vocal sounds were presented rapidly (every 300 ms), it mimicked the continuous flow of auditory information one can experience in real life. The voices presented come from variable unfamiliar speakers differing in age and sex, thus avoiding any familiarity confound in tagging voice selectivity.[17] In fact, in most developmental studies investigating voice processing, the stimuli are often very specific

exemplars—such as lullabies,[23] sentences with "falling intonation contour,"[60] or recordings in motherese (the particular way in which mothers speak to their infants, characterized by "unique acoustic signature" and "exaggerated pitch contour"), [24,25] or only a single female speaker.[27] Contrasting voices that are acoustically very distinct from any other non-vocal sounds therefore do not solve the presence of a categorical, high-level response to voices independent from acoustic features.[52] Here, the periodic presentation of highly heterogeneous voices already provides an intrinsic control for the low-level auditory cues that are not common to the variety of voices used in our experiments.[41] In addition, voices are contrasted with a variety of non-vocal sounds, such that several acoustic features are commonly present in both vocal and non-vocal stimuli.[41] To further control for the specific spectral content of the voice stimuli, we presented the infants with sequences created by scrambling vocal and non-vocal sounds. This technique preserves the spectral content of the original sound and the temporal structure, while altering recognizability and harmonicity.[7,40] Although the selective response at the voice presentation frequency was significant for both intact and scrambled sequences, six channels (three in each hemisphere) were significant for the intact voices as opposed to only two isolated channels for the scrambled sequence (one per hemisphere). Interestingly, these channels do not overlap between intact and scrambled sequences, suggesting different neural sources. In particular, while the voice-selective response to intact voices is descriptively larger in the left than the right hemisphere, no significant lateralization has been found, much like the same response in adults with a similar paradigm.[7]

When directly contrasting the brain responses elicited by intact and scrambled voices, a clear advantage is found for intact voices at three left temporal and two right occipito-temporal channels, whereas no channels respond preferentially to the scrambled voices. A stronger amplitude for intact than scrambled voices is found even when controlling for the number of channels to compare and for the difference between the two conditions for the general auditory response. The results were confirmed using cluster-based permutation testing and by an additional classification analysis over the whole scalp. This reliable enhanced response to intact over scrambled voices reveals that the frequency content and the envelope of the vocal sounds cannot alone explain the preferential brain response to voices, suggesting instead a high-level categorization process. These results observed in young infants of 4 months show important similarities with adult data for which the voice-selective response peaked over the temporal channels even when sounds were further controlled for harmonicity-to-noise ratio.[7] Importantly, we identified individual voice-selective channels and found that a preferential response for intact rather than scrambled voices was reliably present in most of the infants tested, confirming the consistency and the robustness of the voice-selective response in the 4-month-old infant brain. The fact that we were able to unequivocally show voice selectivity, even at an individual level, is rather unique in developmental studies and likely originates in the sensitivity of the FPAS-EEG approach.

It is important to note that our scrambling procedure directly controls for specific features of the sound, such as spectral content and envelope, while not controlling for others, such as harmonicity. The scrambling procedure disrupts the harmonicity of the sounds, which may thus contribute to the difference between the intact and scrambled conditions. However, there are several reasons to consider that the voice-selective response is not fully feature-based and partly reflects a high-level categorical process that goes beyond low-level acoustic cues to categorize variable stimuli as voices. First, because we used a large set of naturalistic auditory stimuli for both objects and voices, low-level acoustic features are highly variable across stimuli. The use of widely variable stimuli controls for irrelevant properties by variability rather than elimination or homogenization, as imposed in natural conditions. This guarantees that the variable physical properties are less prone to elicit a differential neural response that generalizes across category exemplars. In fact, the voice presented may be from a female or a male speaker, young or old, with different emotional content, distinguished from a horn, keychain, bark of a dog, etc. A response to voices is thus present only if, in each 1.11 Hz cycle, there is a specific and reliable neural signature despite the high physical variability of vocal and non-vocal sounds. In other words, the signal at 1.11 Hz and harmonic in the EEG spectrum only captures what generalizes across the neural responses to the 33 voice excerpts within a sequence, an invariant voice representation in the brain.

Second, in the adult study that we based the current study on,[7] another condition was added wherein vocal sounds were embedded together with musical instruments, and all stimuli were controlled for harmonicity, pitch, and spectral center of gravity. Even in this condition, a strong voice-selective response was present at the temporal electrodes, revealing that these features alone do not drive the adult response. Admittedly, adult data should be considered with caution when interpreting infant data. Future infant studies should thus use the same condition, together with other scrambling procedures (e.g., spectrally rotated stimuli that preserve temporal structure but alter spectral information[61]) to further delineate the mechanisms subtending the selective response to voices in the infant brain. Our study, being the first to implement FPAS and frequency-tagging EEG to investigate voice processing in the infant brain, paves the way for many future studies addressing the early development of auditory categorization.

Previous studies on voice processing in young infants mostly focused on preferential response to the mother's or father's voice,[17,62] which are not conclusive about sensitivity to voices as a specific category per se because the responses observed may be triggered by the perception of a familiar sound.[21] Another developmental study focused on the discrimination of affective voices,[27] where emotionally spoken syllables from one female speaker were contrasted with matched synthesized non-vocal sounds. However, without naturalistic auditory objects contrasted with the voices, their discrimination may be driven by naturalness or familiarity rather than a real ability to categorize voices against non-vocal sounds belonging to other categories that the infant encounters in its environment.[27] A previous study showed that when various voices are contrasted with non-vocal sounds, a selective response to voice stimuli has been suggested in 7- but not in 4-month-old infants, therefore suggesting that voice specialization only arises between 4 and 7 months of age.[29] Additional discrepancies regarding the onset of voice selectivity during development come from studies showing a

greater activation for vocal contrasted with non-vocal stimuli in the temporal cortex of infants ranging from 3 to 7 months of age in fMRI[26] and 4 to 7 months of age using fNIRS.[28] Not only can the wide age range used not solve the debate on the ontogeny of voice selectivity but these studies also lack a control for the low-level acoustic features that characterize voices to disentangle their role in high-level voice categorization. To the limit of our knowledge, our study is the first to demonstrate a robust categorical response to voices as early as 4 months of age, where voices are not only highly variable and contrasted with a variety of non-vocal objects but vocal sounds are also controlled for important low-level acoustic features. The fact that we found a voice-selective response as early as 4 months when other studies did not[29] is likely due to the high sensitivity of the FPAS approach, which could become an important tool for developmental auditory neuroscience.

In conclusion, our study demonstrates that 4-month-old infants already possess the neural wiring to automatically categorize voices against many other non-vocal sounds, with a voice-selective response distributed over specific temporal electrodes resembling the topography observed in adults.[7] Taken together, these results bring forth evidence that the infant brain is endowed with the ability to rapidly develop a functionally selective response to voices, a neural selectivity that may support the efficient processing of this socially relevant category of sounds early in life.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Stimuli
  - Design
  - Procedure
  - EEG recording and preprocessing
  - Frequency-domain analysis
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.11.042.

ACKNOWLEDGMENTS

This work was supported in part by a European Research Council starting grant (MADVIS grant #337573) to O.C., the Belgian Excellence of Science program (EOS Project no. 30991544) to O.C., the FLAG-ERA HBP PINT-MULTI (R.8008.19) to O.C., a Mandat d'impulsion scientifique from F.R.S.-FNRS to O.C., a Bourse de Séjour Scientifique from F.R.S.-FNRS to R.P.C., and the French National Research Agency (grant ANR-19-CE28-0009) to A.L. R.C.P. and F.M.B. are research fellows, and O.C. is a senior researcher at the F.R.S.-FNRS. The authors would like to thank Véronique Boulanger for her logistical support in the recruitment of the families, the parents, and their infants who took part in the study; Giada Lettieri for commenting on an earlier version of the manuscript; and Luca Cecchetti for precious input during the reviewing process.

AUTHOR CONTRIBUTIONS

Conceptualization, R.P.C., D.R., F.M.B., A.K., S.T., A.L., and O.C.; methodology, R.P.C., F.M.B., S.T., and O.C.; software, R.P.C., D.R., and F.M.B.; formal analysis, R.P.C.; investigation, R.P.C., D.R., and A.K.; resources, R.P.C., A.L., and O.C.; data curation, R.P.C., D.R., and A.K.; writing – original draft, R.P.C.; writing – review & editing, R.P.C., D.R., F.M.B., S.T., A.L., and O.C.; visualization, R.P.C., D.R., and A.L.; supervision, A.L. and O.C.; project administrations, A.L. and O.C.; funding acquisition, R.P.C., A.L., and O.C.

DECLARATION OF INTERESTS

The authors declare no conflict of interests.

Received: April 18, 2023
Revised: October 20, 2023
Accepted: November 20, 2023
Published: December 13, 2023

REFERENCES

1. Belin, P., Zatorre, R.J., and Ahad, P. (2002). Human temporal-lobe response to vocal sounds. Brain Res. Cogn. Brain Res. *13*, 17–26.
2. Belin, P., Fecteau, S., and Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. Trends Cogn. Sci. *8*, 129–135.
3. Frühholz, S., and Belin, P. (2018). The Oxford Handbook of Voice Perception (Oxford University Press).
4. Warren, J.D., Scott, S.K., Price, C.J., and Griffiths, T.D. (2006). Human brain mechanisms for the early analysis of voices. NeuroImage *31*, 1389–1397.
5. Wenndt, S.J. (2016). Human recognition of familiar voices. J. Acoust. Soc. Am. *140*, 1172.
6. Agus, T.R., Thorpe, S.J., Suied, C., and Pressnitzer, D. (2010). Characteristics of human voice processing. Proceedings of the 2010 IEEE International Symposium on Circuits and Systems (IEEE), pp. 509–512.
7. Barbero, F.M., Calce, R.P., Talwar, S., Rossion, B., and Collignon, O. (2021). Fast periodic auditory stimulation reveals a robust categorical response to voices in the human brain. eNeuro *8*. 0471–20.2021.
8. Suied, C., Agus, T.R., Thorpe, S.J., Mesgarani, N., and Pressnitzer, D. (2014). Auditory gist: recognition of very short sounds from timbre cues. J. Acoust. Soc. Am. *135*, 1380–1391.
9. Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature *403*, 309–312.
10. Agus, T.R., Paquette, S., Suied, C., Pressnitzer, D., and Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. Sci. Rep. *7*, 11526.
11. Trapeau, R., Thoret, E., and Belin, P. (2022). The Temporal Voice Areas are not "just" Speech Areas. Front. Neurosci. *16*, 1075288.
12. Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N.K. (2008). A voice region in the monkey brain. Nat. Neurosci. *11*, 367–374.
13. Andics, A., Gácsi, M., Faragó, T., Kis, A., and Miklósi, A. (2014). Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. Curr. Biol. *24*, 574–578.
14. Belin, P. (2006). Voice processing in human and non-human primates. Philos. Trans. R. Soc. Lond. B Biol. Sci. *361*, 2091–2107.
15. Kisilevsky, B.S., Hains, S.M.J., Lee, K., Xie, X., Huang, H., Ye, H.H., Zhang, K., and Wang, Z. (2003). Effects of experience on fetal voice recognition. Psychol. Sci. *14*, 220–224.

16. Ockleford, E.M., Vince, M.A., Layton, C., and Reader, M.R. (1988). Responses of neonates to parents' and others' voices. Early Hum. Dev. *18*, 27–36.

17. Beauchemin, M., González-Frankenberger, B., Tremblay, J., Vannasing, P., Martínez-Montes, E., Belin, P., Béland, R., Francoeur, D., Carceller, A.M., Wallois, F., et al. (2011). Mother and stranger: an electrophysiological study of voice processing in newborns. Cereb. Cortex *21*, 1705–1711.

18. Purhonen, M., Kilpeläinen-Lees, R., Valkonen-Korhonen, M., Karhu, J., and Lehtonen, J. (2005). Four-month-old infants process own mother's voice faster than unfamiliar voices—electrical signs of sensitization in infant brain. Brain Res. Cogn. Brain Res. *24*, 627–633.

19. Dehaene-Lambertz, G., Montavont, A., Jobert, A., Allirol, L., Dubois, J., Hertz-Pannier, L., and Dehaene, S. (2010). Language or music, mother or Mozart? Structural and environmental influences on infants' language networks. Brain Lang. *114*, 53–65.

20. Lee, G.Y., and Kisilevsky, B.S. (2014). Fetuses respond to father's voice but prefer mother's voice after birth. Dev. Psychobiol. *56*, 1–11.

21. Grosbras, M.-H., and Belin, P. (2020). Development of voice perception in the human brain. In The Social Brain: A Developmental Perspective, J. Decety, ed.

22. Plante-Hébert, J., Boucher, V.J., and Jemel, B. (2021). The processing of intimately familiar and unfamiliar voices: specific neural responses of speaker recognition and identification. PLoS ONE *16*, e0250214.

23. Wild, C.J., Linke, A.C., Zubiaurre-Elorza, L., Herzmann, C., Duffy, H., Han, V.K., Lee, D.S.C., and Cusack, R. (2017). Adult-like processing of naturalistic sounds in auditory cortex by 3- and 9-month old infants. NeuroImage *157*, 623–634.

24. Adam-Darque, A., Pittet, M.P., Grouiller, F., Rihs, T.A., Leuchter, R.H.-V., Lazeyras, F., Michel, C.M., and Hüppi, P.S. (2020). Neural correlates of voice perception in newborns and the influence of preterm birth. Cereb. Cortex *30*, 5717–5730.

25. Hepper, P.G., Scott, D., and Shahidullah, S. (1993). Newborn and fetal response to maternal voice. J. Reprod. Infant Psychol. *11*, 147–153.

26. Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., Deeley, Q., Barker, G.J., Renvall, V., Deoni, S., et al. (2011). Early specialization for voice and emotion processing in the infant brain. Curr. Biol. *21*, 1220–1224.

27. Cheng, Y., Lee, S.Y., Chen, H.Y., Wang, P.Y., and Decety, J. (2012). Voice and emotion processing in the human neonatal brain. J. Cogn. Neurosci. *24*, 1411–1419.

28. Lloyd-Fox, S., Blasi, A., Mercure, E., Elwell, C.E., and Johnson, M.H. (2012). The emergence of cerebral specialization for the human voice over the first months of life. Soc. Neurosci. *7*, 317–330.

29. Grossmann, T., Oberecker, R., Koch, S.P., and Friederici, A.D. (2010). The developmental origins of voice processing in the human brain. Neuron *65*, 852–858.

30. de Heering, A., and Rossion, B. (2015). Rapid categorization of natural face images in the infant right hemisphere. eLife *4*, e06564.

31. Graven, S.N., and Browne, J.V. (2008). Auditory development in the fetus and infant. Newborn Infant Nurs. Rev. *8*, 187–193.

32. Hyde, D.C., Flom, R., and Porter, C.L. (2016). Behavioral and neural foundations of multisensory face-voice perception in infancy. Dev. Neuropsychol. *41*, 273–292.

33. Stets, M., Stahl, D., and Reid, V.M. (2012). A meta-analysis investigating factors underlying attrition rates in infant ERP studies. Dev. Neuropsychol. *37*, 226–252.

34. Geangu, E., Quadrelli, E., Lewis, J.W., Macchi Cassia, V., and Turati, C. (2015). By the sound of it. An ERP investigation of human action sound processing in 7-month-old infants. Dev. Cogn. Neurosci. *12*, 134–144.

35. Hervé, E., Mento, G., Desnous, B., and François, C. (2022). Challenges and new perspectives of developmental cognitive EEG studies. NeuroImage *260*, 119508.

36. Noreika, V., Georgieva, S., Wass, S., and Leong, V. (2020). 14 challenges and their solutions for conducting social neuroscience and longitudinal EEG research with infants. Infant Behav. Dev. *58*, 101393.

37. Levy, D.A., Granot, R., and Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. Psychophysiology *40*, 291–305.

38. Adrian, E.D., and Matthews, B.H.C. (1934). The interpretation of potential waves in the cortex. J. Physiol. *81*, 440–471.

39. Regan, D. (1989). Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine. Clin. Neurophysiol. *119*, 731–743.

40. Dormal, G., Pelland, M., Rezk, M., Yakobov, E., Lepore, F., and Collignon, O. (2018). Functional preference for object sounds and voices in the brain of early blind and sighted individuals. J. Cogn. Neurosci. *30*, 86–106.

41. Rossion, B., Torfs, K., Jacques, C., and Liu-Shuang, J. (2015). Fast periodic presentation of natural images reveals a robust face-selective electrophysiological response in the human brain. J. Vision *15*, 15.1.18.

42. Lochy, A., Van Belle, G., and Rossion, B. (2015). A robust index of lexical representation in the left occipito-temporal cortex as evidenced by EEG responses to fast periodic visual stimulation. Neuropsychologia *66*, 18–31.

43. Jacques, C., Retter, T.L., and Rossion, B. (2016). A single glance at natural face images generate larger and qualitatively different category-selective spatio-temporal signatures than other ecologically-relevant categories in the human brain. Neuroimage *137*, 21–33.

44. Aslin, R.N. (2007). What's in a look? Dev. Sci. *10*, 48–53.

45. Firestone, C. (2020). Performance vs. competence in human–machine comparisons. Proc. Natl. Acad. Sci. USA *117*, 26562–26571.

46. Leleu, A., Rekow, D., Poncet, F., Schaal, B., Durand, K., Rossion, B., and Baudouin, J.Y. (2020). Maternal odor shapes rapid face categorization in the infant brain. Dev. Sci. *23*, e12877.

47. Rekow, D., Leleu, A., Poncet, F., Damon, F., Rossion, B., Durand, K., Schaal, B., and Baudouin, J.-Y. (2020). Categorization of objects and faces in the infant brain and its sensitivity to maternal odor: further evidence for the role of intersensory congruency in perceptual development. Cogn. Dev. *55*, 100930.

48. Rekow, D., Baudouin, J.Y., Poncet, F., Damon, F., Durand, K., Schaal, B., Rossion, B., and Leleu, A. (2021). Odor-driven face-like categorization in the human infant brain. Proc. Natl. Acad. Sci. USA *118*, e2014979118.

49. Norcia, A.M., Appelbaum, L.G., Ales, J.M., Cottereau, B.R., and Rossion, B. (2015). The steady-state visual evoked potential in vision research: a review. J. Vision *15*, 4.

50. Blumberg, M.S., and Adolph, K.E. (2023). Protracted development of motor cortex constrains rich interpretations of infant cognition. Trends Cogn. Sci. *27*, 233–245.

51. Giordano, B.L., McAdams, S., Zatorre, R.J., Kriegeskorte, N., and Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. Cereb. Cortex *23*, 2025–2037.

52. Leaver, A.M., and Rauschecker, J.P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. *30*, 7604–7612.

53. Lewis, J.W., Talkington, W.J., Walker, N.A., Spirou, G.A., Jajosky, A., Frum, C., and Brefczynski-Lewis, J.A. (2009). Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. J. Neurosci. *29*, 2283–2296.

54. Cirelli, L.K., Spinelli, C., Nozaradan, S., and Trainor, L.J. (2016). Measuring neural entrainment to beat and meter in infants: effects of music background. Front. Neurosci. *10*, 229.

55. Lenc, T., Peter, V., Hooper, C., Keller, P.E., Burnham, D., and Nozaradan, S. (2023). Infants show enhanced neural responses to musical meter frequencies beyond low-level features. Dev. Sci. *26*, e13353.

56. Chen, A., Peter, V., and Burnham, D. (2016). Auditory ERP response to successive stimuli in infancy. PeerJ *4*, e1580.

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

58. Adibpour, P., Lebenberg, J., Kabdebon, C., Dehaene-Lambertz, G., and Dubois, J. (2020). Anatomo-functional correlates of auditory development in infancy. Dev. Cogn. Neurosci. *42*, 100752.

59. Morgan, S.T., Hansen, J.C., and Hillyard, S.A. (1996). Selective attention to stimulus location modulates the steady-state visual evoked potential. Proc. Natl. Acad. Sci. USA *93*, 4770–4774.

60. Moon, C., Zernzach, R.C., and Kuhl, P.K. (2015). Mothers Say "baby" and their newborns do not choose to listen: a behavioral preference study to compare with ERP results. Front. Hum. Neurosci. *9*, 153.

61. Green, T., Rosen, S., Faulkner, A., and Paterson, R. (2013). Adaptation to spectrally-rotated speech. J. Acoust. Soc. Am. *134*, 1369–1377.

62. DeCasper, A.J., and Fifer, W.P. (1980). Of human bonding: newborns prefer their mothers' voices. Science *208*, 1174–1176.

63. Retter, T.L., and Rossion, B. (2016). Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream. Neuropsychologia *91*, 9–28.

64. Volfart, A., Rice, G.E., Lambon Ralph, M.A., and Rossion, B. (2021). Implicit, automatic semantic word categorisation in the left occipito-temporal cortex as revealed by fast periodic visual stimulation. Neuroimage *238*, 118228.

65. Fujioka, T., Mourad, N., He, C., and Trainor, L.J. (2011). Comparison of artifact correction methods for infant EEG applied to extraction of event-related potential signals. Clin. Neurophysiol. *122*, 43–51.

66. Mourad, N., Reilly, J.P., De Bruin, H., Hasey, G., and MacCrimmon, D. (2007). A simple and fast algorithm for automatic suppression of high-amplitude artifacts in EEG data. In IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP, *2007*, pp. 1–393.

67. Talwar, S., Barbero, F.M., Calce, R.P., and Collignon, O. (2023). Automatic brain categorization of discrete auditory emotion expressions. Brain Topogr. *36*, 854–869.

68. Luck, S.J. (2014). An Introduction to the Event-Related Potential Technique, second edition (MIT Press).

69. Retter, T.L., Rossion, B., and Schiltz, C. (2021). Harmonic amplitude summation for frequency-tagging analysis. J. Cogn. Neurosci. *33*, 2372–2393.

70. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

71. Quek, G.L., Liu-Shuang, J., Goffaux, V., and Rossion, B. (2018). Ultra-coarse, single-glance human face detection in a dynamic visual stream. NeuroImage *176*, 465–476.

72. Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

73. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Original Data	This paper	https://osf.io/cvhk3/
Software and Algorithms		
63 Ag/AgCl electrode cap, Waveguard	ANT Neuro, The Netherlands	https://www.ant-neuro.com/products/waveguard_original
MATLAB_R2016b	MathWorks, USA	RRID:SCR_001622
RStudio	RStudio Team, 2018	RRID:SCR_013715
Letswave6 toolbox	https://www.letswave.org	https://github.com/NOCIONS/letswave6

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Collignon Olivier (olivier.collignon@uclouvain.be)

Materials availability

The study did not produce new materials.

Data and code availability

Original data reported in this paper is available at https://osf.io/cvhk3. This paper does not report original code.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Twenty-three 4-month-old infants took part in the experiment (12 females, mean age = 125.8 ± 13 (SD) days, range: 107–164 days). All participants were born at term, had an APGAR score at 9/10 at birth, and their caregivers reported no auditory or neurological disorders. No participants' data was excluded from the final sample. The size was estimated a priori as in Rekow et al.[48] (i.e., Cohen's d = 0.8, significance level α = 0.05 (two-tailed), power 1-β = 0.95; we chose a high-power value to improve our ability to find a significant voice-selective response). Caregivers were recruited through the local birth registry after a mail contact, and they gave written consent after being fully informed about the aim and procedure of the study. The experiment was conducted according to the Declaration of Helsinki and under the approval of the French ethics committee (CPP Sud-Est III).

METHOD DETAILS

Stimuli

Stimuli were the same as in a previous study on voice categorization in adults.[7] The total set consisted of 192 sounds divided in 137 non-vocal sounds (environmental sounds, musical instruments, sounds produced by objects) and 55 vocal sounds (speech and non-speech vocalizations by speakers of different sex, age, and emotional states) that were equalized in overall energy (root mean square).[51] Each sound lasted 250 ms, including a fade-in and a fade-out of 10 ms each to avoid clicking. Importantly, each sound was available in both intact and scrambled versions. Scrambled sounds were created following a procedure based on a previous study[7,40] by applying a Fast Fourier Transformation to all original sounds, and by shuffling the magnitude and phase of the signal within each 200 Hz frequency window, as a wider window would lead to recognizable sounds. An inverse Fourier Transform was then computed, and the original sound envelope, which was extracted using Hilbert's transform, was applied to the scrambled sound. As a result, scrambled sounds had frequency content and spectral-temporal structure almost identical to those of the original sounds [Figure 1], but harmonicity was altered and recognizability disrupted.[7] Moreover, the magnitude of the FFT of the envelope of intact and scrambled sequences are similar ($t_{(19)}$= -1.1920, p= 0.2479, Cohen's d = -0.1751). To facilitate individual sound segregation and avoid masking effect within the rapid stimulation sequence, a 50-ms silence was inserted after each sound, thus resulting in 300-ms-long sound stimuli.

Design

The intact and scrambled sounds were respectively embedded to create the corresponding two conditions. Each stimulation sequence was created by randomly selecting 106 stimuli from the whole set of sounds (71 non-vocal and 35 vocal sounds), resulting

in unique sequences at each stimulation. The sounds used to create the sequences were systematically presented in a random order. Therefore, each participant listened to unique exemplars of intact and scrambled sequences in the effort to increase generalization. The total duration of a sequence was 31.80 s, including a 1-s fade-in and a 1-s fade-out at the beginning and at the end of the sequence, respectively. Sequences were structured so that sounds were presented at a rate of 3.33 Hz (due to the 300-ms stimulus duration, i.e., 1/0.300 s; see section Stimuli). Crucially, vocal sounds were inserted at each third sound, leading to an additional rate of voice presentation at 1.11 Hz (3.33 Hz/3 = 1.11 Hz) leading to a cycle of 900ms (object – object – voice; OOV) (Figure 1D). With this design, a response at the general rate of sound presentation (3.33 Hz) and its harmonics (i.e., integer multiples) for both intact and scrambled conditions would indicate the infant brain's ability to segment all presented sounds. Most importantly, a response at the 1.11-Hz voice presentation frequency and its harmonics would reflect the infant brain's ability to discriminate vocal from non-vocal sounds and to generalize this response to the heterogenous vocal sounds presented within a sequence. Any advantage for the intact compared to the scrambled condition would demonstrate that the voice-selective response is not simply driven by the frequency content and the envelope of the sounds.

Procedure

After electrode-cap placement, infants were installed on a seat in a light- and sound-attenuated room. The seat was positioned at a 60 cm distance in front of a monitor behind which speakers were centered. Sequences were presented at an intensity of 60 dB, measured with a decibel meter at the distance of the seat from the speakers. Caregivers were asked to stay distant from their infants during stimulation and not to interact with them unless in case of manifest distress. To maximize testing duration and the number of trials presented, when infants could not continue the experiment in the seat, they were further tested on their caregiver's lap, who was instructed to remain silent, to not move nor interact with them. Infants were continuously monitored using a camera placed on top of the monitor and non-periodic visual distractors were used when needed to reorient attention toward the speakers. Stimulation sequences were launched by the experimenter when the infant was still and at least 5 seconds after the end of the previous one. A sequence was a priori rejected and not considered for analyses in case of excessive movement or manifest distress before the end of the sequence. Only one sequence was discarded for one infant. To equalize the number of sequences per condition, intact and scrambled sequences were presented in a pseudo-randomized order (i.e., the order of presentation was shuffled at each paired condition presentation so that the same condition was not presented more than two times in a row) until the first sign of tiredness, discomfort, or caregiver's demand. Excluding the discarded sequence, infants performed between 7 and 16 sequences in each condition (total range: 14–32 minutes of testing, mean ± SD 22.04 min ± 4.99 min sequences, for an overall testing duration per infant between 7.46 min and 17.06 min.

EEG recording and preprocessing

EEG was continuously recorded from a 63 Ag/AgCl electrode cap (Waveguard, ANT Neuro, The Netherlands) according to the 10-10 classification system with a sampling rate of 1000 Hz (acquisition reference: CPz, ground: AFz, the magnitude of initial electrode impedance < 30 kΩ). Data analysis was performed using the Letswave6 toolbox (https://www.letswave.org) running on MATLAB_R2016b (MathWorks, USA), and custom-build scripts in RStudio (RStudio Team, 2018). Analysis followed a procedure validated in adults (for auditory, FPAS[7] or visual studies[63,64]) and infants.[30,46,47] Raw EEG data were bandpass filtered at 0.1 – 100 Hz (Butterworth filter, 4^{th} order) and down-sampled to 200 Hz to reduce file size. Individual EEG data was then cropped according to each sequence segment starting from 2 seconds before the fade-in until the end of the sequence. Artifacts were corrected in each sequence using the Artifact Blocking Algorithm[65,66] with a threshold of ± 250 μV. All 63 channels were then re-referenced according to a common average reference. Individual preprocessed data were re-segmented starting at the first complete cycle of voice presentation after the fade-in (i.e.,1.2 s after the start of the sequence) until the end of the sequence to contain an integer number of cycles avoiding overspill of the response in the frequency domain. The resulting sequences had a length of 30.6 s (34 cycles of 900 ms, see Design) and were averaged separately by participant and condition (intact and scrambled) in the time domain, resulting in a single averaged 30.6-s epoch per condition and participant. This step allows to increase signal-to-noise ratio (SNR) by attenuating the EEG activity that is not in phase with the auditory stimulation.

Frequency-domain analysis

Frequency-domain analysis was then performed to isolate and quantify both the general auditory (3.33 Hz) and the voice-selective (1.11 Hz) responses and their harmonics (i.e., integer multiples). A Fast Fourier Transformation (FFT) was first applied, resulting in individual amplitude spectra for each channel and condition. Amplitude spectra had a resolution of $\approx$ 0.0327 Hz (1/30.6 s; total bins: 6120), allowing a high-resolution noise level estimation from frequency bins surrounding the signal at the frequency of interest. The number of harmonics to consider for further analysis was then determined. For each condition, data were grand-averaged across infants and channels, resulting in a single amplitude spectrum per condition. Z-scores were then calculated by subtracting the mean amplitude of the noise from the amplitude of the signal at the frequency bins of interest and then dividing by the standard deviation of the noise. Noise was estimated from 24 surrounding bins, excluding the local maximum and minimum amplitude and the two immediate adjacent bins.[63] The number of surrounding bins was chosen according to standard practices in the field of visual categorization in adults[63] and auditory categorization in adults.[7,67] This is a compromise between having enough bins to estimate background noise but still having a ''local'' estimation as the EEG frequency spectrum is differently affected by noise at different frequencies.[68] Significant harmonics were determined for each neural response separately, excluding harmonics of the general

auditory response (e.g., 3.33 Hz) for the voice-selective response. Harmonics were considered as significant if their z-score was higher than 1.64 ($p < .05$, one-tailed, signal > noise), and were included for further analysis until two consecutive non-significant z-scores were found in either one of the two conditions. The highest number of harmonics in any condition was used to define the number of harmonics for the two conditions. Following empirical recommendation,[69] the decision to include consecutive harmonics in the analysis followed the rationale that including a high number of, even non-significant, harmonics to estimate the overall response is a better option than missing a part of the response, especially considering that the overall magnitude of the response is not influenced by non-significant harmonics (i.e., with a baseline-corrected amplitude ≈ 0, see below).[69] The resulting number of harmonics was seven for the general auditory responses (from 3.33 to 23.31 Hz) and two for the voice-selective responses (1.11 Hz and 2.22 Hz) (Table S1). For the following analyses, chunks centered on each harmonic and surrounded by 12 frequency bins on each side were cropped from individual amplitude spectra for each infant, channel, condition, and neural response (7 chunks for the general auditory response, 2 chunks for the voice-selective response), and summed across harmonics. Henceforth, if not stated otherwise, the general auditory response and voice-selective response will refer to these summed responses across significant harmonics. The magnitudes and signal-to-noise ratios (SNR) of each of these responses were then calculated. For the former, a baseline correction was computed to quantify each response in a single value in microvolts irrespective of the noise. The mean amplitude of the noise (same estimation as for z-scores, see above) was subtracted from the amplitude of the signal, leading to a baseline-corrected amplitude ≈ 0 in the absence of response. To calculate the SNR, the amplitude of the signal was divided by the mean amplitude of the noise (as defined above), leading to a SNR of 1 in the absence of response (signal = noise) and a SNR > 1 in case of a response. Baseline-corrected amplitudes were used for both statistical and illustration purposes, while SNR was only used for illustration.

Statistical analysis

For each response, we first determined its significance over the whole scalp using the grand-averaged response across infants for each condition. Z-scores (as defined in section Frequency-domain analysis) were computed at each electrode to compare the amplitude of the signal and surrounding noise. P-values were corrected across electrodes using the False Discovery Rate (FDR) procedure to control for multiple comparisons[70] ($p < .05$, FDR-corrected for 63 electrodes, one-tailed, signal > noise). Next, to investigate whether the general auditory and voice-selective responses differ between intact and scrambled conditions, grand-averaged responses in the scrambled condition were subtracted from grand-averaged responses in the intact condition, and Z-scores were calculated at each channel ($p < .05$, FDR-corrected for 63 electrodes, two-tailed, intact $\neq$ scrambled). Given the high signal-to-noise ratio, we also computed the same analysis at an individual level. For each subject, the response in the scrambled condition was subtracted from the response in the intact condition and Z-scores were calculated at each channel ($p < .05$, FDR-corrected for 63 electrodes, two-tailed, intact $\neq$ scrambled).

Moreover, for the voice-selective response, we also compared its baseline-corrected amplitude between intact and scrambled conditions considering the variance across infants. We first considered the mean baseline-corrected amplitude across the channels previously identified as showing a significant response in each condition and compared them using a two-tailed paired t-test ($p < .05$, intact $\neq$ scrambled). Additionally, to further investigate this main effect, we analyzed the magnitude of the response separately for each region, given that the voice-selective response is distributed over left temporal and right occipito-temporal scalp regions (see in result section, voice-selective response). We averaged the baseline-corrected amplitude of two channels per region, namely the one with the highest Z-scores identified in the intact condition and in the scrambled condition. We followed this procedure to ensure a fair comparison across conditions, having the same numbers of channels from each condition to obtain the mean amplitudes to compare. The lateralization of the voice–selective response was also investigated by computing a lateralization index that estimates the size of hemispheric asymmetry reported to the overall response across both hemispheres. For each infant, the uncorrected response amplitude of the three significant channels identified in each hemisphere for the standard condition was averaged, obtaining one value per hemisphere. Next, the difference between the left-hemispheric response and the right-hemispheric response was computed and divided by the sum of the two responses, expressed as: (left – right) / (left + right). The obtained index represents hemispheric asymmetry expressed as a percentage, with positive and negative values indicating right- and left-lateralized responses, respectively. Significance of this index was assessed using a t-test against 0 ($p < .05$, two-tailed).

In addition, we determined whether the difference between the two conditions remains consistent when the interindividual variability is considered, being the voice-selective responses focal across individual scalp locations. In each subject, we selected the channels with the highest magnitude (one in the intact and one in the scrambled condition) independently of the location, following the individual channel-of-interest procedure.[71] An additional t-test was then conducted between the mean of the individual channels across conditions. We also conducted two additional analyses to reinforce our finding that intact voices elicit a stronger selective response than scrambled voices. First, a nonparametric cluster-based permutation testing[72] was applied with t-tests between the summed responses for each condition ($p < .05$, two-tailed, intact $\neq$ scrambled) calculated at each electrode for cluster size definition and percentiles from the distribution of mean cluster sum obtained after 1000 permutations as cluster significance threshold ($p < .05$, two-tailed). To reinforce the idea that the intact and scrambled conditions elicit different brain responses in the infant's brain, we used a decoding approach to determine whether conditions can be classified from their topographical distribution. We performed a binary classification (using SVM – support vector machine classifier[73]) using a leave-one-trial-out cross-validation scheme. The baseline-corrected-amplitude response of one trial from both conditions were used for testing while all other epochs were used for training in a k-fold cross validation strategy, where k was the number of epochs in any condition for each participant. Individual

accuracies of each fold were then averaged to obtain a single accuracy for each subject. Statistical significance of the averaged binary classification accuracy was assessed using permutation testing (1000 permutations) against a null distribution.

Lastly, to ensure that any advantage of the intact condition for the voice-selective response is not accounted for by the same advantage for the general auditory response, the voice-selective response was normalized by the general auditory response in each condition. Raw and baseline-corrected amplitudes of the voice-selective response were divided by the mean scalp-wide baseline-corrected amplitude of the general auditory response in each condition separately and analyses were computed following the same procedure as above.