

Emergence of neural encoding of auditory objects while listening to competing speakers

Nai Ding^a and Jonathan Z. Simon^{a,b,1}

Departments of ^aElectrical and Computer Engineering and ^bBiology, University of Maryland, College Park, MD 20742

Edited by Thomas D. Albright, The Salk Institute for Biological Studies, La Jolla, CA, and approved June 6, 2012 (received for review April 5, 2012)

A visual scene is perceived in terms of visual objects. Similar ideas have been proposed for the analogous case of auditory scene analysis, although their hypothesized neural underpinnings have not yet been established. Here, we address this question by recording from subjects selectively listening to one of two competing speakers, either of different or the same sex, using magnetoencephalography. Individual neural representations are seen for the speech of the two speakers, with each being selectively phase locked to the rhythm of the corresponding speech stream and from which can be exclusively reconstructed the temporal envelope of that speech stream. The neural representation of the attended speech dominates responses (with latency near 100 ms) in posterior auditory cortex. Furthermore, when the intensity of the attended and background speakers is separately varied over an 8-dB range, the neural representation of the attended speech adapts only to the intensity of that speaker but not to the intensity of the background speaker, suggesting an object-level intensity gain control. In summary, these results indicate that concurrent auditory objects, even if spectrotemporally overlapping and not resolvable at the auditory periphery, are neurally encoded individually in auditory cortex and emerge as fundamental representational units for top-down attentional modulation and bottom-up neural adaptation.

spectrotemporal response function | reverse correlation | phase locking | selective attention

In a complex auditory scene, humans and other animal species can perceptually detect and recognize individual auditory objects (i.e., the sound arising from a single source), even if strongly overlapping acoustically with sounds from other sources. To accomplish this remarkably difficult task, it has been hypothesized that the auditory system first decomposes the complex auditory scene into separate acoustic features and then binds the features, as appropriate, into auditory objects (1–4). The neural representations of auditory objects, each the collective representation of all the features belonging to the same auditory object, have been hypothesized to emerge in auditory cortex to become fundamental units for high-level cognitive processing (5–7). The process of parsing an auditory scene into auditory objects is computationally complex and cannot as yet be emulated by computer algorithms (8), but it occurs reliably, and often effortlessly, in the human auditory system. For example, in the classic “cocktail party problem,” where multiple speakers are talking at the same time (9), human listeners can selectively attend to a chosen target speaker, even if the competing speakers are acoustically more salient (e.g., louder) or perceptually very similar (such as of the same sex) (10).

To demonstrate an object-based neural representation that could subserve the robust perception of an auditory object, several key pieces of evidence are needed. The first is to demonstrate neural activity that exclusively represents a single auditory object (4, 7). In particular, such an object-specific representation must be demonstrated in a range of auditory scenes with reliable perception of that auditory object, and especially in challenging scenarios in which the auditory object cannot be easily segregated by any basic acoustic features, such as frequency or binaural cues. For this reason, we investigate the existence of object-specific auditory representations by using an auditory scene consisting of a pair of

concurrent speech streams mixed into a single acoustic channel. In this scenario, the two speech streams each form a distinct perceptual auditory object but they overlap strongly in time and frequency, and are not separable using spatial cues. Therefore, any neural representation of an auditory object (i.e., in this case, a single stream of speech) would not emerge without complex segregation and grouping processes.

Second, the neural processing of an auditory object must also be adaptive and independent (2, 4). In particular, the neural processing of each auditory object should be modulated based on its own behavioral importance and acoustic properties, without being influenced by the properties of other auditory objects or the stimulus as a whole. Building on the well-established phenomena of feature-based top-down attentional modulation (11–14) and feature-based bottom-up neural adaptation to sound intensity (15), we investigate here whether such top-down and bottom-up modulations occur separately for individual auditory objects (i.e., in an object-based manner). Specifically, using this speech segregation paradigm, we ask the listeners to attend to one of the two speakers while manipulating separately the intensity of the attended and background speakers. If an observed neural representation is object-based, not only must it be enhanced by top-down attention but it must adapt to the intensity change of that speech stream alone, without being affected by the intensity change of the other stream or of the mixture as a whole.

In this study, we investigate whether a robust neural representation of an auditory object can be observed in the brain, and when and where it might emerge. In the experiment, the subjects selectively listened to one of two concurrent spoken narratives mixed into a single acoustic channel, answering comprehension questions about the attended spoken narrative after each 1-min stimulus. The neural recordings were obtained using magnetoencephalography (MEG), which is well suited to measure spatially coherent neural activity synchronized to speech rhythms (i.e., the slow temporal modulations that define the speech envelope) (16–19). Such spatially coherent phase-locked activity is strongly modulated by attention (20–22) and has been hypothesized to play a critical role in grouping acoustic features into auditory objects (3).

Specifically, we hypothesize that in cortical areas with an object-based representation, neural activity should phase lock to the rhythm of a single auditory object, whereas in cortical areas in which object-based representations are not yet formed, or formed only weakly, the neural response should phase lock to the envelope of the entire physical stimulus (i.e., the speech mixture) (both examples are illustrated in Fig. 1*A* and *B*). In other words, what is encoded in the neural response, whether a single speech stream or the mixture, can be easily distinguished by which sound's

Author contributions: N.D. and J.Z.S. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹To whom correspondence should be addressed. E-mail: jzsimon@umd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1205381109/-DCSupplemental.

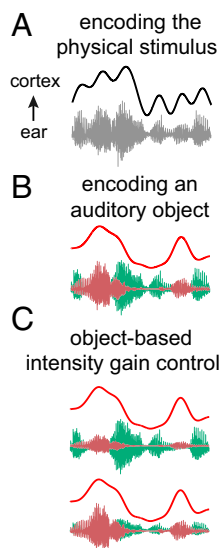


Fig. 1. Illustration of object-based neural representations. Here, the auditory scene is illustrated using a mixture of two concurrent speech streams. (A) If a complex auditory scene is not neurally parsed into separate auditory objects, cortical activity (*Upper*, curve) phase locks to the temporal envelope of the physical stimulus [i.e., the acoustic mixture (*Lower*, waveform)]. (B) In contrast, using the identical stimulus (but illustrated here with the unmixed instances of speech in different colors), for a hypothetical neural representation of an individual auditory object, neural activity would instead selectively phase lock to the temporal envelope only of that auditory object. (C) Neural representation of an auditory object should, furthermore, neurally adapt to an intensity change of its own object (*Upper*) but should remain insensitive to intensity changes in another auditory object (*Lower*). Neither of these modifications to the acoustic stimulus therefore significantly changes the neural representation (comparing A and C).

rhythm it is synchronized to. Critically, bottom-up neural adaptation to sound intensity is also investigated. Neural adaptation also determines whether a neural representation is object-based based or not, depending on which sound stream (or mixture) the neural representation adapts to. We do this by analyzing the phase-locked neural activity when the intensity of the attended speaker and the background speaker is manipulated separately (Fig. 1C). These hypothesized, object-specific neural representations are investigated and revealed, using single-trial neural recordings and an advanced neural decoding method that parallels state-of-the-art analysis methods used in functional MRI (fMRI) (23) and intracranial recording (24, 25).

Results

Deciphering the Spatial-Temporal Code for Individual Speakers. In the first experiment, listeners selectively listened to one of two competing speakers of different sex, mixed into a single acoustic channel with equal intensity. To probe object-specific neural representations, we reconstructed the temporal envelope of each of the two simultaneous speech streams by optimally integrating MEG activity over time and space (i.e., sensors). Such a reconstruction of the envelope of each speech stream, rather than the physical stimulus, can be successful only if the stimulus mixture is neurally segregated (“unmixed”) and the speech streams of the two speakers are represented differentially. We first reconstructed the temporal envelope of the attended speech. Fig. 2A shows representative segments of the different envelopes reconstructed by this decoder, from listeners hearing the identical speech mixture but attending to different speakers in it. Clearly, the reconstructed envelope depends strongly on the attentional focus of the listener and resembles the envelope of the attended speech. At the single-subject level and the single-trial level, the reconstructed envelope is

more strongly correlated with the envelope of the attended speaker than of the unattended speaker ($P < 0.001$, paired permutation test; Fig. 2B, *Left*). This attention-dependent neural reconstruction is seen in 92% of trials (Fig. S1).

We also reconstructed the temporal envelope of the background speech using a second decoder that integrates neural activity spatiotemporally in a different way. The result of this reconstruction is indeed more correlated with the envelope of the background speech rather than of the attended speech ($P < 0.005$, paired permutation test; Fig. 2B, *Right*). Therefore, by integrating the temporal and spatial neural responses in two distinct ways, the attended and background speech can be successfully decoded separately. On average, the reconstruction for the background speech is more correlated with the background speech in 73% of the trials from individual subjects (Fig. S1; significantly above chance level; $P < 0.002$, binomial test). In this experiment, the speakers are of opposite sex, but the neural representations of segregated speech streams can be similarly demonstrated even for

Neural Encoding of Each Speech Stream

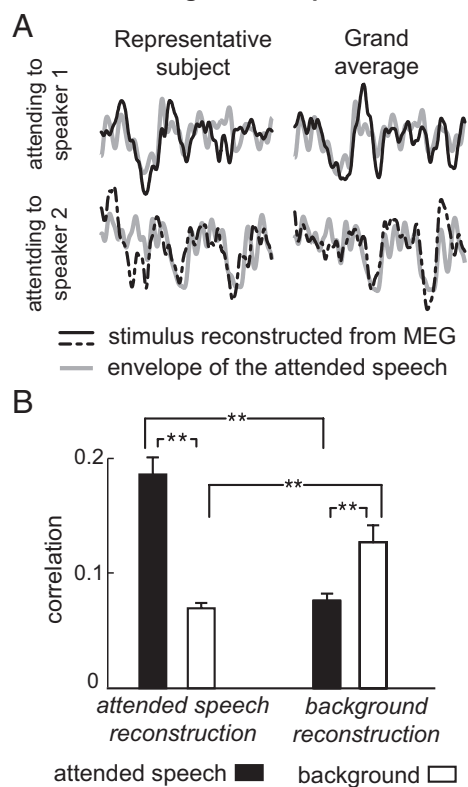


Fig. 2. Decoding the cortical representation specific to each speech stream. (A) Examples of the envelope reconstructed from neural activity (black), superimposed on the actual envelope of the attended speech when presented in isolation (gray). (*Upper* and *Lower*) Different envelopes are decoded from neural responses to identical stimuli, depending on whether the listener attends to one or the other speaker in the speech mixture, with each resembling the envelope of the attended speech. Here, the signals, 5 s in duration, are averaged over three trials for illustrative purposes, but all results in the study are based on single-trial analysis. (B) Two separate decoders reconstruct the envelope of the attended and background speech, respectively, from their separate spatial-temporal neural responses to the speech mixture. The correlation between the decoded envelope and the actual envelope of each speech stream is shown in the bar graph (averaged over trials and speakers), with each error bar denoting 1 SEM across subjects ($**P < 0.005$, paired permutation test). The separate envelopes reconstructed by the two decoders selectively resemble that of attended and background speech, demonstrating a separate neural code for each speech stream.

the more challenging scenario where the two speakers are of the same sex (Fig. S2). Consequently, these results demonstrate that the neural representation in auditory cortex goes beyond encoding just the physically presented stimulus (the speech mixture) and shows selective phase locking to auditory objects (Fig. S3).

Robustness to the Intensity of Either Speaker. When the intensity of either of the two competing speaker changes, up to 10 dB, human listeners can still understand either speaker with more than 50% intelligibility (10). Intensity gain control may contribute to this robustness in speech perception. Here, we address whether intensity gain control occurs globally for an auditory scene or separately for each auditory object. In a second Varying-Loudness experiment, the intensity level of one speech stream, either the attended or the background, is kept constant, whereas the other is reduced (up to 8 dB). Under this manipulation, the intensity ratio between the attended and background speakers [i.e., the target-to-masker ratio (TMR)] ranges between -8 dB and 8 dB.

To distinguish how different intensity gain control mechanisms would affect the neural representation of each speech stream, we simulate possible decoding outcomes (SI Methods). The MEG activity is simulated by the sum of activity precisely phase locked to each speech stream and interfering stimulus-irrelevant background activity. The strength of the phase-locked activity is normalized by either the strength of whole stimulus, for a global gain control mechanism, or the strength of the encoding auditory object, for an object-based gain control mechanism. The simulated decoding outcomes under different gain control mechanisms are shown in Fig. 3A.

The neural decoding from actual MEG measurements is shown in Fig. 3B. For the decoding of the attended speech, the decoded envelope is significantly more correlated with the envelope of the attended speech [$P < 0.004$, $F(1,71) = 25.8$, attentional focus \times TMR two-way repeated measures ANOVA], and this correlation is not affected by TMR. The result is consistent with the object-based gain control model rather than with the global gain control model. Similarly, the neural decoding of the background speech is also affected by the attentional focus [$P < 0.02$, $F(1,71) = 14.65$, higher correlation with the background speech, two-way ANOVA], without interaction between attention and TMR. Consequently, the neural representation of a speech stream is stable both against the intensity change of that stream and against the intensity change of the other stream, consistent with the hypothesized object-specific gain control (compare the examples shown in Fig. 1C).

Spatial Spectrotemporal Response Function and Neural Sources. The decoding analysis above integrates neural activity, spatiotemporally, to reveal an object-specific neural representation optimally. To characterize the neural code that the decoder extracts information from, we analyze the neural encoding process via the spectrotemporal response function (STRF) for each MEG sensor (26, 27). The linear STRF and the linear decoder are, respectively, the forward and backward models describing the same relationship between the stimulus and neural response. Nevertheless, only the forward STRF model can reveal the timing and spatial information of the neural encoding process.

The STRF functionally describes how the spectrotemporal acoustic features of speech are transformed into cortical responses. It deconvolves the neural activity evoked by the continuous envelope of speech. In this STRF model, the encoding of each speech stream is modeled using the auditory spectrogram (28) of the unimixed speech signal with unit intensity. The STRF shows the neural response to sound features at different acoustic (i.e., carrier) frequencies (as labeled by the y axis). At any given frequency, the horizontal cross-section of the STRF (e.g., Fig. 4A) characterizes the time course of the neural response evoked by a unit power increase of the stimulus at that frequency for one MEG sensor.

Neural Adaptation to the Intensity of Each Speech Stream

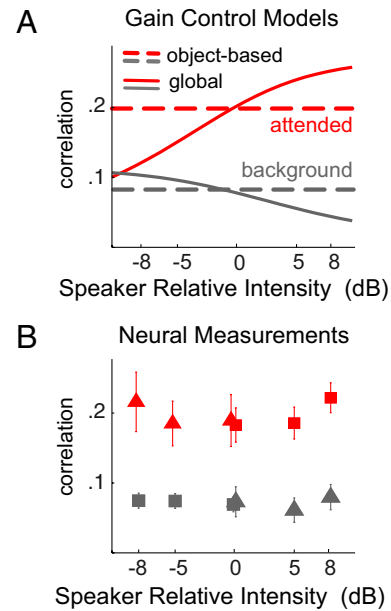


Fig. 3. Decoding the attended speech over a wide range of relative intensity between speakers. (A) Decoding results simulated using different gain control models. The x axis shows the intensity of the attended speaker relative to the intensity of the background speaker. The red and gray curves show the simulated decoding results for the attended and background speakers, respectively. Object-based intensity gain control predicts a speaker intensity invariant neural representation, whereas the global gain control mechanism does not. (B) Neural decoding results in the Varying-Loudness experiment. The cortical representation of the target speaker (red symbols) is insensitive to the relative intensity of the target speaker. The acoustic envelope reconstructed from cortical activity is much more correlated with the attended speech (red symbols) than the background speech (gray symbols). Triangles and squares are results from the two speakers, respectively.

The MEG STRF contains two major response components (Fig. S4): one with latency near 50 ms, here called the $M50_{STRF}$, and the other with latency near 100 ms, here called the $M100_{STRF}$. This indicates that two major neural response components continuously follow the temporal envelope of speech, with delays of 50 ms and 100 ms, respectively. Because an STRF is derived for each MEG sensor, the neural source locations of the $M50_{STRF}$ and $M100_{STRF}$ can be estimated based on the distribution over all sensors of the strength of each component (i.e., the topography of the magnetic fields at each latency). An equivalent current dipole (ECD)-based neural source analysis reveals that the $M50_{STRF}$ and $M100_{STRF}$ responses arise from different neural sources (SI Methods), with the source of $M100_{STRF}$ being 5.5 ± 1.5 mm and 7.1 ± 2.0 mm more posterior in the left and right hemispheres, respectively (Fig. 4B; $P < 0.005$ for both hemispheres, paired t test). The ECD location of the neural source of the $M100_{STRF}$ peak is consistent with that observed for the M100 response to tone pips, localized to the superior temporal gyrus and roughly in the planum temporale (29).

The amplitudes of the $M50_{STRF}$ and $M100_{STRF}$ are further analyzed in the neural source space, based on the STRF at the ECD location of each component. The amplitude of the $M100_{STRF}$ is much stronger for the attended speech than for the background speech [Fig. 4A; $P < 0.007$, $F(1,87) = 11.85$, attentional focus \times hemisphere \times speaker, three-way repeated-measures ANOVA]. The amplitude of the $M50_{STRF}$ is, in contrast, not significantly modulated by either attention or TMR. The latency of the $M50_{STRF}$

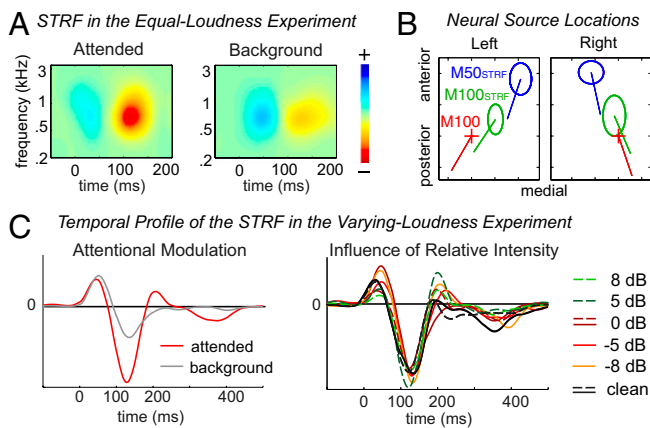


Fig. 4. Cortical encoding of the spectral-temporal features of different speech streams. (A) STRFs for the attended and background speech, at the neural source location of the M100_{STRF}. Attention strongly enhances the response with latency near 100 ms. (B) Neural source locations for the M50_{STRF} and M100_{STRF} in each hemisphere, as estimated by dipole fitting. The location of the neural source of the M50_{STRF} is anterior and medial to that of the M100_{STRF} and M100. The source location for each subject is aligned based on the source of the M100 response to tone pips, shown by the cross. The span of each ellipse is 2 SEM across subjects. The line from each dipole location illustrates the grand averaged orientation of each dipole. Each tick represents 5 mm. (C) Temporal profile of the STRF in the Varying-Loudness experiment for the attended speech. The M100_{STRF} (averaged over TMR) is strongly modulated by attention, whereas the M50_{STRF} is not (Left). Neither response peak is affected by the intensity change of the two speakers (Right).

and M100_{STRF} is also modulated by attention [$P < 0.03$, $F(1,187) > 7$ for both peaks, three-way repeated-measures ANOVA] and is, on average, 11 and 13 ms shorter, respectively, when attended.

The temporal profile of the STRF in the Varying-Loudness experiment is shown in Fig. 4C, which is extracted by applying singular value decomposition to the STRF. The M100_{STRF} is significantly modulated by attention [$P < 0.03$, $F(1,143) = 9.4$, attentional focus \times hemisphere \times speaker \times TMR, four-way repeated-measures ANOVA], whereas the M50_{STRF} is not. Neither response component is affected by TMR (compare the examples shown in Fig. 1C). The invariance of the M50_{STRF} and M100_{STRF} to the intensity of both the attended and background speech streams provides further evidence for the hypothesized object-specific gain control.

Discussion

This study investigates whether a multisource auditory scene, perceptually represented in terms of auditory objects, is neurally represented in terms of auditory objects as well. From subjects selectively listening to one of two spectrotemporally overlapping speech streams, we do observe neural activity selectively synchronized to the speech of a single speaker (as is illustrated in Fig. 1B). Furthermore, in an ecologically valid listening setting, this selective representation of an individual speech stream is both modulated by top-down attention and normalized by the intensity of that sound stream alone (as is illustrated in Fig. 1C). In sum, this meets all the criteria of an object-based representation, [e.g., those specified by Griffiths and Warren (4)]: The observed neural representation is selective to the sound from a single physical source, is minimally affected by competing sound sources, and is insensitive to perceptually unimportant acoustic variations of the stimulus (e.g., changes in intensity).

Temporal Coherence, Attention, and Object-Based Representations.

The object-specific representations seen here are precisely synchronized to the temporal envelope of speech. In speech and

natural sounds in general, the temporal envelope synchronizes various acoustic features, including pitch and formant structures. Therefore, they provide important cues for perceptual auditory grouping (30) and are critical for robust speech recognition. For example, major speech segregation cues, such as pitch, are not sufficient for speech recognition, whereas acoustic features necessary for speech recognition (e.g., the spectrotemporal envelope) are not easily distinguishable between speakers. A solution to this dilemma would be to group acoustic features belonging to the same auditory object, both speech segregation and intelligibility-relevant cues, through temporal coherence analysis, and then to process selectively the attended auditory object as a whole (3). In other words, the auditory cortex selects the attended speech stream by amplifying neural activity synchronized to the coherent acoustic variations of speech (i.e., the envelope). This idea is highly consistent with the large-scale synchronized and object-specific activity seen in this study.

At the neuronal mechanistic level, it is plausible that the low-frequency, phase-locked neural activity binds features belonging to the same object by regulating the excitability of neurons, such that a given neural network will be more responsive when processing features from the corresponding auditory object (22). Furthermore, such a rhythmic regulation of neuronal excitability may also contribute to the segmentation of continuous speech into perceptual units (e.g., syllables) (16).

Hierarchical Processing of Auditory Objects in Auditory Cortex. Of the two major neural response components that track the speech envelope, the M100_{STRF} is much more strongly modulated by attention than the M50_{STRF}. These two neural response components track the speech envelope with different latencies and are generated from distinct neural sources. Based on their positions relative to the neural source of the M100 (29), the centers of neuronal current generating the M50_{STRF} and M100_{STRF} are dominantly from Heschl's gyrus and the planum temporale, respectively (31). The latency and source location of the two components demonstrate a hierarchy of auditory processing (32, 33), and the representation of the attended object becomes dominant from shorter to longer latency activity and from core to posterior auditory cortex. Therefore, although auditory object representations may start to emerge as early as primary auditory cortex (7), the top-down attentional modulation of the large-scale, object-based neural representation may emerge most strongly with later processing.

The routing of the neural processing of the attended auditory object into posterior auditory cortex may generally underlie the selection of auditory information when there are competing spectrotemporally complex auditory objects. MEG studies have shown that selectively listening to sound embedded in a complex auditory scene modulates longer latency (~100–250 ms) responses in association auditory cortex but not the shorter latency (~50 ms) steady-state response in core auditory cortex (21, 34, 35), as is also the case for MEG/EEG responses to transient sounds (13, 36–38). PET studies also indicate that the areas posterior to core auditory cortex are more activated when speech is interfered by temporally modulated noise than stationary noise (39, 40), because modulated noise contains speech-like features and requires additional processes of information selection. Furthermore, a recent fMRI study has shown that in a multitalker environment, the planum temporale is increasingly activated when the number of information sources (i.e., speakers) increases (41). Taken together, these results support the idea that posterior auditory cortex plays a major role in the generation of auditory objects (42, 43) and the selection of information based on the listener's interest (25, 44).

Neural Adaptation to the Intensity of Individual Auditory Object. The recognition of speech relies on the shape of its spectrotemporal modulations and not its mean intensity. This study demonstrates that cortical activity is precisely phase locked to the temporal

modulations but insensitive to the mean intensity of the speech streams, and therefore is effectively encoding only the shape of the modulations. Intensity gain control has been demonstrated in multiple stages of the auditory system (15) and constitutes an auditory example of neural normalization, which has been suggested as a canonical neural computation (45). For example, in the cochlear nucleus, neurons encode the shape of the spectral modulation of speech (e.g., a vowel) invariantly to its mean intensity (46).

Critically different from these previous studies, however, the encoding of temporal modulations seen here is invariant to the intensity of each speech stream rather than the overall intensity of the mixture. In the Varying-Loudness experiment, the intensity of one speaker changes, whereas the other is kept constant. Maintaining a stable representation despite the altered speech requires the observed neural adaptation to the sound intensity of the specific speech stream. The stable representation of the constant speaker, in contrast, requires the observed lack of adaptation to the overall intensity of the sound mixture, which covaries with the intensity of the altered speech. These both contrast with the simpler mechanism of global intensity gain control, which would require the neural representation of both speech streams to be modulated in the same way based on the overall intensity of the acoustic stimulus. Therefore, the data strongly suggest the existence of an object-specific intensity gain control, which normalizes the strength of neural activity based on the intensity of individual auditory objects.

In sum, this study demonstrates the key signatures of an object-specific neural representation arising from the analysis of a complex auditory scene. Such an object-specific neural representation is phase locked to the slow rhythms (<10 Hz) of the encoded auditory object, and it adapts to the intensity of that object alone. Under the modulation of top-down attention, the auditory response in posterior auditory cortex (latency near 100 ms) dominantly represents the attended speech, even if the competing speech stream is physically more intense. This object-specific auditory representation provides a bridge between feature-based, precisely phase-locked sensory responses and interference-resilient cognitive processing and recognition of auditory objects.

Methods

Subjects, Stimuli, and Procedure. Twenty normal-hearing, right-handed, young adult native speakers of American English (between 18 and 26 y old) participated in the experiment in total. Eleven (5 female) participated in the Equal-Loudness experiment, 6 (3 female) participated in the Varying-Loudness experiment, and 3 (2 female) participated in the Same-Sex experiment.

The stimuli contain three segments from the book *A Child's History of England* by Charles Dickens, narrated by three different readers (2 female).

Each speech mixture was constructed by mixing the two speakers digitally in a single channel and was then divided into sections with a duration of 1 min. All stimuli were delivered identically to both ears using tube phones plugged into the ear canals. In the Equal-Loudness and Varying-Loudness experiments, the two simultaneous speakers were of opposite sex (mean pitch separation of 5.5 semitones). In the Equal-Loudness experiment, the two speakers were mixed, with equal rms values of sound amplitude. The subjects were instructed to focus on only one speaker and to answer questions related to the comprehension of the passage they focused on. In the Varying-Loudness experiment, the stimuli were mixtures of the same two speakers. The intensity of one speaker was fixed, whereas the intensity of the other speaker was either the same or was -5 or -8 dB weaker. The Same-Sex experiment used mixtures of speakers of the same sex. Details of the three experiments are included in *SI Methods*.

Data Recording and Processing. The neuromagnetic signal was recorded using a 157-channel, whole-head MEG system (KIT) in a magnetically shielded room, with a sampling rate of 1 kHz. The ongoing neural response (excluding the first second) during each 1-min stimulus was filtered between 1 and 8 Hz (21), downsampled to 40 Hz, and was then used for the decoding and STRF analysis.

The temporal envelope of each speaker in the stimulus was decoded by linearly integrating the spatial-temporal brain activity. The decoder was optimized so that the decoded envelope was maximally correlated with the speaker to decode and minimally correlated with the other speaker (*SI Methods*). All correlations in this study were measured by the absolute value of the Pearson's correlation coefficient. The decoder optimized in this way was a discriminative model that reconstructed an envelope similar to one speaker but distinct from the other, and it was therefore used to explore the neural code unique to each speaker. The decoder was applied to individual trials, and the percentage of trials where decoding was successful (decoded envelope being more correlated with the intended speaker) is always reported as the grand average. This decoding approach effectively explores both the spatial and temporal information in MEG and avoids the sometimes ill-posed problem of estimating the neural source locations.

The STRF was used to model how speech features are encoded in the MEG response (21), in contrast to how the decoders transform MEG activity (backward) to speech features. A single STRF transforms the spectrotemporal features of speech to a single-response waveform. Therefore, because of the multichannel nature of MEG, a complete forward model is described as a 3D spatial-STRF model (MEG sensor position \times frequency \times time). The MEG response was modeled as the sum of the responses to individual speakers, with each having unit intensity (more details about STRF estimation are provided in *SI Methods*). The MEG data were averaged over trials in the STRF analysis for each stimulus and attentional condition.

ACKNOWLEDGMENTS. We thank David Poeppel, Mary Howard, Monita Chatterjee, Stephen David, and Shihab Shamma for comments and Max Ehrmann for technical assistance. This research was supported by the National Institute of Deafness and Other Communication Disorders Grant R01-DC-008342.

- Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
- Shinn-Cunningham BG (2008) Object-based auditory and visual attention. *Trends Cogn Sci* 12(5):182–186.
- Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* 34(3):114–123.
- Griffiths TD, Warren JD (2004) What is an auditory object? *Nat Rev Neurosci* 5: 887–892.
- Fishman YI, Steinschneider M (2010) Formation of auditory streams. *The Oxford Handbook of Auditory Science: The Auditory Brain*, eds Rees A, Palmer A (Oxford Univ Press, New York), Vol 2, pp 215–245.
- Snyder JS, Gregg MK, Weintraub DM, Alain C (2012) Attention, awareness, and the perception of auditory scenes. *Front Psychol* 3:15.
- Nelken I, Bar-Yosef O (2008) Neurons and objects: The case of auditory cortex. *Front Neurosci* 2(1):107–113.
- Wang D, Brown GJ (2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE, New York).
- Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am* 25:975–979.
- Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* 109:1101–1109.
- Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA (2009) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* 61:317–329.
- Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci* 6:1216–1223.
- Hillyard SA, Hink RF, Schwent VL, Picton TW (1973) Electrical signs of selective attention in the human brain. *Science* 182:177–180.
- Xiang J, Simon J, Elhilali M (2010) Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration. *J Neurosci* 30:12084–12093.
- Robinson BL, McAlpine D (2009) Gain control mechanisms in the auditory pathway. *Curr Opin Neurobiol* 19:402–407.
- Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54:1001–1010.
- Abrams DA, Nicol T, Zecker S, Kraus N (2008) Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *J Neurosci* 28:3958–3965.
- Ahissar E, et al. (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci USA* 98:13367–13372.
- Koskinen M, et al. (2012) Identifying fragments of natural speech from the listener's MEG signals. *Hum Brain Mapp*, 10.1002/hbm.22004.
- Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J Neurosci* 30:620–628.
- Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol* 107:78–89.
- Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci* 32:9–18.
- Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. *Nature* 452:352–355.

24. Pasley BN, et al. (2012) Reconstructing speech from human auditory cortex. *PLoS Biol* 10:e1001251.
25. Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236.
26. Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J Neurophysiol* 85:1220–1234.
27. deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. *Science* 280:1439–1443.
28. Yang X, Wang K, Shamma SA (1992) Auditory representations of acoustic signals. *IEEE Trans Inf Theory* 38:824–839.
29. Lütkenhöner B, Steinsträter O (1998) High-precision neuromagnetic study of the functional organization of the human auditory cortex. *Audiol Neurootol* 3(2–3):191–213.
30. Sheft S (2007) Envelope processing and sound-source perception. *Auditory Perception of Sound Sources*, eds Yost WA, Popper AN, Fay RR (Springer, New York), pp 233–279.
31. Steinschneider M, Liégeois-Chauvel C, Brugge JF (2011) Auditory evoked potentials and their utility in the assessment of complex sound processing. *The Auditory Cortex*, eds Winer JA, Schreiner CE (Springer, New York), pp 535–559.
32. Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393–402.
33. Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat Neurosci* 12:718–724.
34. Gutschalk A, Micheyl C, Oxenham AJ (2008) Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biol* 6:e138.
35. Okamoto H, Stracke H, Bermudez P, Pantev C (2011) Sound processing hierarchy within human auditory cortex. *J Cogn Neurosci* 23:1855–1863.
36. Woods DL, Hillyard SA, Hansen JC (1984) Event-related brain potentials reveal similar attentional mechanisms during selective listening and shadowing. *J Exp Psychol Hum Percept Perform* 10:761–777.
37. Woldorff MG, et al. (1993) Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proc Natl Acad Sci USA* 90:8722–8726.
38. Ahveninen J, et al. (2011) Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. *Proc Natl Acad Sci USA* 108:4182–4187.
39. Scott SK, Rosen S, Beaman CP, Davis JP, Wise RJS (2009) The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. *J Acoust Soc Am* 125:1737–1743.
40. Scott SK, Rosen S, Wickham L, Wise RJS (2004) A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *J Acoust Soc Am* 115:813–821.
41. Smith KR, Hsieh I-H, Saberi K, Hickok G (2010) Auditory spatial and object processing in the human planum temporale: No evidence for selectivity. *J Cogn Neurosci* 22:632–639.
42. Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. *Trends Neurosci* 25:348–353.
43. Zatorre RJ, Bouffard M, Ahad P, Belin P (2002) Where is 'where' in the human auditory cortex? *Nat Neurosci* 5:905–909.
44. Zion Golumbic EM, Poeppel D, Schroeder CE (2012) Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain & Language*, 10.1016/j.bandl.2011.12.010.
45. Carandini M, Heeger DJ (2012) Normalization as a canonical neural computation. *Nat Rev Neurosci* 13:51–62.
46. Young ED (2008) Neural representation of spectral and temporal information in speech. *Philos Trans R Soc Lond B Biol Sci* 363:923–945.

Supporting Information

Ding and Simon 10.1073/pnas.1205381109

SI Methods

Subjects, Stimuli, and Procedures. All subjects were paid for their participation. The experimental procedures were approved by the University of Maryland Institutional Review Board. Written informed consent form was obtained from each subject before the experiment.

The stimuli consisted of three chapters from a public domain narration of the book *A Child's History of England* by Charles Dickens (<http://librivox.org/a-childs-history-of-england-by-charles-dickens/>). The first was chapter 3, read by a male speaker, and the others were chapter 8 and chapter 22, read by two different female speakers. All speaker pauses (periods of silence longer than 300 ms) were shortened to 300 ms to keep the speech streams flowing continuously, and all stimuli were low-pass-filtered below 4 kHz.

The subjects were required to close their eyes when listening. Before each main experiment, 100 repetitions of a 500-Hz tone pip were presented to each subject to elicit the M100 response, a reliable auditory response measured 100 ms after the onset of a tone pip and whose neural source is easy to localize within auditory cortex (1). Three main experiments were conducted as follows.

Equal-Loudness experiment. The stimulus was a mixture of two spoken narratives with equal rms values. The subjects were instructed to focus on one speaker (speaker 1, male; speaker 2, female) until the stimulus was finished and then to switch focus to the other speaker while the same stimulus was played again. The same process was repeated three times, resulting in three trials with an identical stimulus and attentional focus. Each trial contained two 1-min duration sections. To help the listeners attend to the correct speaker, the first second of each section was replaced by the clean recording from the target speaker. The speaker attended to first was counterbalanced across subjects. After each section, the subjects were asked to answer a question related to the comprehension of the passage they had just attended to. On average, 69% of the questions were correctly answered [not depending on the number of trials; $P > 0.8$, $F(2,32) = 0.2$, one-way repeated measures ANOVA]. After this part of the experiment, the unmixed stimuli (each speaker alone) were presented to the listeners four times. Comprehension questions were interspersed occasionally to ensure the subjects were awake during the whole experiment.

Varying-Loudness experiment. In this experiment, the intensity of speaker 2 was fixed at roughly 75 dB sound pressure level (SPL) and speaker 1 was mixed at either the same intensity, as evaluated by rms value, or at an intensity 5 dB or 8 dB lower. Therefore, when speaker 1 was attended, the attended speaker had constant intensity, whereas the background speaker was reduced. In contrast, when speaker 2 was attended, the background speaker was kept constant, whereas the attended speaker was damped. The target-to-masker ratio (TMR) varied overall from -8 dB to 8 dB. At each TMR, two sections, each 1 min in duration, were presented, after each of which a question was asked. The listeners correctly answered 71% of the questions asked after each minute of listening, which did not significantly vary with TMR ($P > 0.7$, one-way repeated measures ANOVA), indicating that the listeners understood the story without any obvious difficulty, even when the acoustics of stimulus changed dramatically.

The experiment was divided into four blocks, with attention switching from one speaker to the other after each block. For half the subjects, in block 1 the listener focused on speaker 1 and in block 2 the listener focused on speaker 2. In this case the block 1 stimuli were (in presentation order): two sections of clean speech from speaker 1, two sections of speaker 1 mixed with speaker 2

at 8 dB TMR, two sections of speech mixed with 5 dB TMR, and two sections of speech mixed with 0 dB TMR. The block 2 stimuli were (in presentation order): two sections of clean speech from speaker 2, two sections of speaker 2 mixed with speaker 1 at 0 dB TMR, two sections of speech mixed with -5 dB TMR, and two sections of mixed speech mixed with -8 dB TMR. Blocks 3 and 4 were identical to blocks 1 and 2. For the other half of the subjects, the roles of the speakers were switched. The story continued naturally throughout each block. Such an experimental design produces two trials, from alternative blocks, for each section for each attentional condition. Five of the six subjects in this experiment were asked to rate subjectively what percentage of words was correctly recognized after they first listened to each stimulus. The averaged subjective speech intelligibility is 88%, 80%, 68%, 60%, and 48% at 8 dB, 5 dB, 0 dB, -5 dB, and -8 dB TMR, respectively, which varies significantly with TMR [$P < 10^{-4}$, $F(4,24) = 12.6$, one-way repeated measures ANOVA].

Same-Sex experiment. The two chapters read by female speakers were mixed digitally with equal intensity and then divided into six 30-s sections. The average pitch of the two speakers differed by 3.2 semitones (2). The subjects were instructed to focus on one speaker throughout the six sections and then to switch attention to the other speaker when all the sections were played again. This whole process was repeated again, resulting in two trials for each attentional state. To help the subjects identify which speaker to focus on, the first 5 s of each section were replaced by clean speech from the target speaker; therefore, the neural recording of the first 5 s was not included in any analysis. A story comprehension question was asked after each session.

Additionally, each listener went through two initial training sessions before attending to each speaker. In the first session, the nonattended speaker was turned on gradually, using a sigmoidal ramp that saturated after 20 s. The second session used stimuli with the same intensity as the stimuli used in the experiment. The training sessions were repeated on the subjects' request to make sure the subjects were able to identify and focus on the target speaker after the last training session.

Data Recording. A 200-Hz, low-pass filter and a notch filter at 60 Hz were applied to the MEG recording online. Three reference magnetic sensors and three vibrational sensors were used to measure the environmental magnetic field and vibrations, and were also utilized to denoise the magnetoencephalography (MEG) signals (3). Five electromagnetic coils were used to measure each subject's head position inside the MEG machine.

Spectrotemporal Response Function Model. The spectrotemporal representations of the speech streams of the two speakers are denoted as $S_A(f, t)$ and $S_B(f, t)$, respectively, and the MEG response is denoted as $r(t, k)$, where k is an index for MEG sensors. The linear spectrotemporal response function (STRF) model can then be formulated as

$$r(t, k) = \sum_f \sum_\tau STRF_A(f, \tau, k) S_A(f, t - \tau) + \sum_f \sum_\tau STRF_B(f, \tau, k) S_B(f, t - \tau) + \varepsilon(t, k),$$

where $STRF_A(f, t, k)$ and $STRF_B(f, t, k)$ are STRFs for the attended and background speech, respectively, for every MEG sensor and $\varepsilon(t, k)$ is the residual response waveform not explained by the STRF model. The spectrotemporal representations of the speech of the two speakers were calculated from unmixed speech using an auditory filter-bank model (4). The amplitude of the

stimulus, $S_A(f, t)$ and $S_B(f, t)$, is represented in logarithmic amplitude scale (i.e., in decibels). The mean amplitude of the envelope of each speech stream is normalized to 0 dB, because in a linear model like the STRF, the mean of the stimulus is represented by the dc component of the neural response, which is not reliably measurable by MEG. In other words, $S_A(f, t)$ and $S_B(f, t)$ are independent of the actual intensity of either speech stream. Therefore, the amplitude of the STRFs should covary with the intensity of either speech stream, unless the change of stimulus intensity is compensated by the auditory system in an object-based manner.

The STRF model was applied separately to individual sensors. For the sake of computational efficiency, however, the 157-channel MEG dataset was dimensionally reduced to 30 channels when estimating the STRF, using denoising source separation (DSS) (5), but was then transformed back to the MEG sensor space. Each STRF was estimated using boosting with 10-fold cross-validation (6), as described by Ding and Simon (7).

The temporal profile of an STRF is extracted using singular value decomposition (SVD). For the STRF from a MEG sensor or a neural source location, the SVD of STRF is $STRF(f, t) = \sum_p \lambda_p TRF_p(t) STRF_p(f)$. The temporal profile of the STRF, or the temporal response function, is defined as $\lambda_1 TRF_1(t)$ (7).

Extraction of the M50_{STRF} and M100_{STRF} Magnetic Fields. The M50_{STRF} and M100_{STRF} were extracted from two time intervals: 10–100 ms and 50–200 ms, respectively. The approximate latency of each response peak was determined based on a single temporal response function extracted from the spatial STRF using SVD (7). The M100_{STRF}, also known as the M100-like response, is known to have the same polarity as the M100 response evoked by a tone pip (7), whereas the M50_{STRF} has the opposite polarity. Therefore, the M100_{STRF} was determined by the strongest response peak with a magnetic field topography positively correlated with that of the M100, and the M50_{STRF} was determined similarly, but with a negative correlation. The magnetic field pattern extracted for each peak was averaged over speakers and attentional conditions, and it was then used for neural source localization.

Source Space Analysis. In the neural source analysis, subjects from the Equal-Loudness experiment and the Varying-Loudness experiment were pooled together, and the responses at different TMRs were also averaged. The neural source of each peak in the STRF was modeled by an equivalent-current dipole (ECD) in each hemisphere. A spherical head model was derived for each subject using MEG Laboratory software program v.2.001M (Yokogawa Electric; Eagle Technology, Kanazawa Institute of Technology). The position of the ECD was estimated using a global optimization approach (8). The grand averaged correlation between the fitted ECD magnetic field and the measured magnetic field was above 95% in both hemispheres and for both M50_{STRF} and M100_{STRF}. When comparing the ECD positions of different peaks in STRF, we included only ECDs successfully capturing the measured magnetic field, characterized by a higher than 85% correlation between the ECD magnetic field and the measured magnetic field. No more than 2 of the 17 subjects were excluded this way, for each STRF peak. After the ECD positions were determined, the moment of the dipole was estimated using the least squares method (9). The polarity of the M100_{STRF} is defined as negative so as to be consistent with the polarity of the N1 peak of EEG.

In the analysis of the amplitude and latency of the M50_{STRF} and the M100_{STRF}, the STRFs are projected to the lead field of the dipole in each hemisphere. Mathematically, if the STRF is $STRF(f, t, k)$ and the lead field is $L(k)$, the projection is $\sum_k STRF(f, t, k)L(k)$.

Optimal Design of the Decoders. Mathematically, the decoding operation can be formulated as $ENV(t) = \sum_k \sum_s M_k(t + \tau) D_k(\tau)$,

where $ENV(t)$ is the decoded envelope, $M_k(t)$ is the MEG measurement from a single sensor k , and $D_k(t)$ is the linear decoder for the same sensor k .

In the following, we first discuss the case of a single MEG sensor, and therefore drop the index k . In matrix form, the decoding is expressed as $v = Md$, where $v = [ENV(0), ENV(\Delta t), \dots, ENV(T_{MAX})]^T$; $d = [D(0), D(\Delta t), \dots, D(T_D)]^T$; and the matrix, M , is $[M(0), M(0 + \Delta t), \dots, M(0 + T_D); M(\Delta t), M(\Delta t + \Delta t), \dots, M(\Delta t + T_D); \dots; M(T_{MAX}), M(T_{MAX} + \Delta t), \dots, M(T_{MAX} + T_D)]$, with Δt being 25 ms. T_D , the maximal time delay considered by the decoder, is selected to be 500 ms.

Suppose the envelopes of the speech streams of the two speakers are $s_1 = [s_1(0), s_1(\Delta t), \dots, s_1(T_{MAX})]$ and $s_2 = [s_2(0), s_2(\Delta t), \dots, s_2(T_{MAX})]$ and they are normalized to have the same L_2 norm (i.e., $\|s_1\| = \|s_2\|$). The envelope was extracted by summing, over frequency, the spectrotemporal representation of the speech (4), with its amplitude expressed in logarithmic scale. The correlations between the decoded envelope and the envelopes of the two speech streams are $c_1 = \alpha^{-1} s_1^T v = \alpha^{-1} s_1^T M d$ and $c_2 = \alpha^{-1} s_2^T M d$, respectively, where $\alpha^{-1} = \|s_1\| \cdot \|v\| = \|s_2\| \cdot \|v\|$. Let us denote $r_1 = s_1 M$ and $r_2 = s_2 M$; then, $(c_1/c_2)^2 = (d r_1 r_1^T d^T) / (d r_2 r_2^T d^T)$. Denote $R_1 = r_1 r_1^T$ and $R_2 = r_2 r_2^T$; then, it is known that the quantity $(c_1/c_2)^2$ is maximized when d is the generalized eigenvector of R_1 and R_2 with the largest eigenvalue (10).

The conclusion from this single MEG sensor case is easily generalized to the case for multiple MEG sensors by concatenating the recording from all the MEG sensors. For example, in the case of 100 MEG sensors, the first row of M becomes $[M_1(0), M_1(0 + \Delta t), \dots, M_1(0 + T_D), M_2(0), M_2(0 + \Delta t), \dots, M_2(0 + T_D), \dots, M_{100}(0), M_{100}(0 + \Delta t), \dots, M_{100}(0 + T_D)]$ after concatenation. In this study, to reduce the computational complexity, the 157 MEG sensors were compressed into 3 virtual sensors using DSS in each hemisphere (5). Therefore, the 6 virtual sensors were concatenated first; the two covariance matrices, R_1 and R_2 , were then calculated; and, finally, the decoder was obtained by generalized eigen-decomposition.

The chance level performance of the decoders was simulated by independently shuffling the order of each 1-min stimulus (independently between the 2 simultaneous speakers) and the order of all the responses 4,096 times. At this chance level, obtained by reconstructing the stimulus based on unmatched responses, the reconstructed envelope is similarly correlated with the speech envelopes of both speakers ($P > 0.8$, paired t test) and the 95th percentile of the correlation with each speech stream envelope is below 0.01, showing that the decoder does not show bias toward either speaker.

In the Varying-Loudness experiment, the same decoder was used to decode the stimulus at every TMR. The stimulus and response in every TMR condition were divided into a training set and a testing set. All training sets were then pooled together to train the decoder. After training, the decoder was applied to individual TMR conditions to assess the neural encoding accuracy. Therefore, if the decoding results were consistent over TMR conditions, it would imply that the underlying spatial-temporal neural code captured by the decoder is unaffected by the intensity change of a speaker. In Fig. 3, the decoding accuracy for each speech stream is normalized separately. Specifically, the decoding accuracy for one speech stream, the first or the second, is divided by the decoding accuracy of that speech stream when presented individually, c_{s1} or c_{s2} , and is then multiplied by the mean accuracy of decoding a speech stream presented individually [i.e., $(c_{s1} + c_{s2})/2$].

Cross-validation was used to evaluate the performance of decoders: Half of the data in each experimental condition were used to train the decoder, and the other half were used to calculate the correlations between the decoder output and the stimulus envelopes.

Models of Gain Control. The envelope of speech is assumed to be faithfully encoded in auditory cortex, and the imperfect decoding of speech envelope is assumed to be attributable to (stimulus-irrelevant) neural background activity. Therefore, the MEG measurement is modeled as a linear mixture of neural activity phase locked to each speech stream and stimulus-irrelevant spontaneous activity. To simplify the simulation, but without loss of generality, we further assume that the neural encoding of each stream instantaneously follows that speech stream; for example, $r_{A0}(t) = s_A(t)$ and $r_{B0}(t) = s_B(t)$, where $r_{A0}(t)$ and $r_{B0}(t)$ are the raw neural responses to the attended speech and background speech, respectively, and $s_A(t)$ and $s_B(t)$ are the corresponding speech envelopes.

We model the intensity gain control of neural activity using two models. One model normalizes the MEG activity by the strength (measured by the rms) of the envelope of the acoustic mixture [i.e., $s_{\text{mix}}(t)$], and the second model normalizes the MEG activity by the strength of each speaker individually. The two models are described as follows:

i) Global gain control model:

$$r_A(t) = r_{A0}(t)/\text{RMS}(s_{\text{mix}}(t))$$

$$r_B(t) = r_{B0}(t)/\text{RMS}(s_{\text{mix}}(t))$$

ii) Object-based gain control model:

$$r_A(t) = r_{A0}(t)/\text{RMS}(s_A(t))$$

$$r_B(t) = r_{B0}(t)/\text{RMS}(s_B(t))$$

The neural reconstruction of the attended speech, a linear combination of MEG activity, is modeled as $\hat{s}_A(t) = r_A(t) + \lambda_B r_B(t) + \lambda_N n(t)$. In the reconstruction, $r_B(t)$ and $n(t)$ are attenuated but not eliminated because of, for example, the limited spatial resolution of MEG. The two free parameters λ_B and λ_N are fit based on the Equal-Loudness experiment; that is, when $s_A(t)$ and $s_B(t)$ have equal intensity, λ_B and λ_N are adjusted so that the simulated decoding results [i.e., the correlation between $r_A(t)$ and $s_A(t)$ and the correlation between $r_A(t)$ and $s_B(t)$] match the experimental observations in the Equal-Loudness experiment (Fig. 2B). The model is then used to predict the decoding results in the Varying-Loudness experiment, where the intensity of two speakers is changed separately. The model predictions are generally insensitive to the values of λ_B and λ_N .

SI Discussion

Decoding in Delta and Theta Bands. In this decoding analysis, the neural response is filtered between 1 and 8 Hz, including both the delta (1–4 Hz) and theta (4–8 Hz) bands. If only the delta or theta band is analyzed, the decoder output is still more similar to the attended speech ($P < 0.005$, paired t test, for both frequency bands and both speakers). The decoding result is more accurate in the delta band than in the theta band (higher correlation with the attended speaker in 91% vs. 76% of trials).

Segregation of Speakers of the Same Sex. In the Equal-Loudness and Varying-Loudness experiments, the two competing speakers are of different sex. Male and female voices differ in pitch range and formant frequencies and, as basic perceptual categories, are easy to distinguish. Human listeners can also successfully segregate speakers of the same sex, even if they are acoustically and perceptually highly similar to each other. To test if the neural segregation of speakers of the same sex also occurs in auditory cortex, we created a mixture of two female speakers and instructed the subjects to focus on one of them. In this Same-Sex experiment, after a training session, the subjects can successfully recognize and follow the target speaker, and answered 74% of the comprehension questions asked during the experiment. From the neural response, the temporal envelope of the speech of the attended speaker is decoded (Fig. S2), and the decoded envelope

is more correlated with the attended speech than the background speech ($P < 0.01$ for both speakers, paired t test based on individual trials from individual listeners).

The STRF analysis is also applied to the Same-Sex experiment; the amplitude of the $M100_{\text{STRF}}$ is significantly modulated by attention [attentional focus \times hemisphere \times speaker, three-way repeated-measures ANOVA, $P < 0.02$, $F(1,23) = 66$], whereas the amplitude of the $M50_{\text{STRF}}$ is not.

Phase Locking to the Speech Mixture. The two decoders shown in Fig. 2 reconstruct a temporal envelope that is similar to one speech stream but distinct from the other. To distinguish the encoding of either speech stream further not just from the other but from the physically presented speech mixture, we designed an additional pair of decoders. This pair also reconstructs temporal envelopes that are maximally correlated with the envelopes of the attended and background speech, respectively, but are minimally correlated with the envelope of the mixture.

This decoding analysis is also successful. The extracted temporal envelope is more closely correlated with the envelope of the individual speech streams than the temporal envelope of the mixture (Fig. S3; $P < 0.0005$, paired permutation test). Therefore, this pair of decoders directly demonstrates the existence of phase locking to individual speech streams, distinct from the phase locking to the mixture.

In the current experiment, the auditory scene consists of only two auditory objects and we demonstrate that the attended object and the background object are represented differentially. For the case of more than two auditory objects in an auditory scene, whether the neural system divides the scene into multiple objects or only the attended object and the background must be determined by future experiments.

Speech Segmentation and Intelligibility. Phase-locked neural activity, such as seen here, encodes the temporal information of speech and is a candidate for use in the process of segmentation of continuous speech into syllabic-level structures (11). Speech recognition, however, further requires the decoding of phonemic content, which itself requires spectral analysis. Therefore, the correct segmentation of speech is necessary but not sufficient for intelligibility. For example, it has been shown that the ability to use syllabic stress cues to segment speech is maintained even for unintelligible speech at -15 dB TMR (12). In the Varying-Loudness experiment of the current study, the subjects could generally understand the story at all tested TMRs but reported a TMR-dependent word recognition rate, ranging from 50 to 90%, consistent with the literature (13, 14). In contrast, the neural tracking of the temporal features of speech is not significantly affected by TMR, and therefore is not correlated with the word recognition rate. It is thus likely that the phase-locked neural activity observed by MEG is related to the temporal processing and syllabic-level segmentation of speech rather than directly to the much higher order phenomenon of speech intelligibility (11, 15, 16).

In the Varying-Loudness experiment, the subjectively rated speech intelligibility decreases with TMR, consistent with objectively measured speech intelligibility (13). This close match between subjectively rated and objectively measured speech intelligibility has been demonstrated previously (17). The percentage of questions correctly answered, however, is not significantly affected by TMR. It is possible that this lack of effect is attributable to the small number of questions ($n = 4$) being asked for each TMR condition. The purpose of these questions, however, is only to ensure the subjects' attention. The fact that the subjects can correctly answer most questions in every TMR condition demonstrates the subjects' engagement in the task and their ability to understand the story roughly, at every tested TMR.

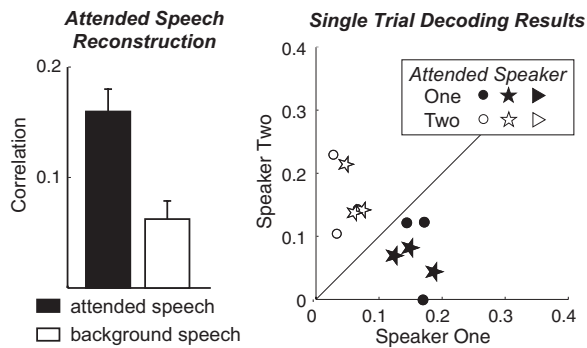


Fig. 52. Decoding of the speech representations for two competing female speakers. The correlation between the decoded envelope and the actual envelope is shown by the bar graph (averaged over trials and subjects) and the scatterplot (each trial and subject separately). The attended speech can be decoded exclusively from the neural response to the mixture.

Distinguishing the Neural Encoding of Each Speech Stream and the Neural Encoding of Speech Mixture

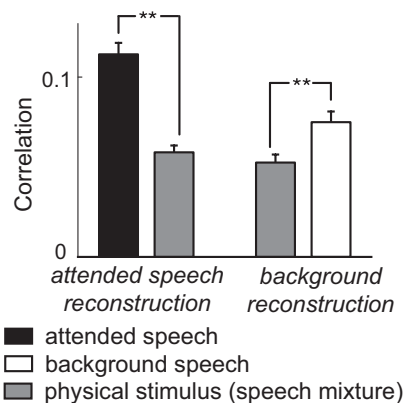


Fig. 53. Distinguishing the encoding of each speaker with the encoding of the temporal envelope of the speech mixture. Two decoders are designed to extract the temporal envelopes of the attended and background speech, respectively, while minimizing the correlation between the decoder output and the envelope of the mixture. The two decoders' outputs are indeed more correlated with the envelopes of individual speakers than with the envelope of the mixture ($*P < 0.0005$).

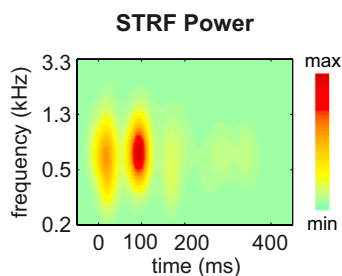


Fig. 54. STRF power is shown as a function of frequency and time (summed over all sensors and subjects) for unmixed speech. It is dominated by two response components, $M50_{STRF}$ and $M100_{STRF}$, with respective latencies near 50 ms and 100 ms, respectively.