Article

Neuron

# Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception

## Highlights

- Mixed speech is encoded differently in primary and nonprimary auditory cortex (AC)
- Primary AC selectively represented individual speakers unchanged with attention
- Nonprimary AC represented the attended speaker invariant to acoustic overlap
- These results show the neural underpinnings of auditory object formation in AC

## Authors

James O'Sullivan, Jose Herrero, Elliot Smith, ..., Sameer A. Sheth, Ashesh D. Mehta, Nima Mesgarani

## Correspondence

nima@ee.columbia.edu

## In Brief

How different areas of the human auditory cortex (AC) represent mixed speech is unclear. O'Sullivan et al. obtained invasive recordings from subjects listening to multi-talker speech. They found that the primary AC represented the individual speakers and was unchanged by attention. In contrast, the nonprimary AC selectively represented the attended speaker, was invariant to the acoustic overlap with unattended speaker, and was linearly predictable from the primary AC. These results reveal the neural underpinnings of the hierarchical formation of auditory objects in human AC.

O'Sullivan et al., 2019, Neuron *104*, 1195–1209
December 18, 2019 
https://doi.org/10.1016/j.neuron.2019.09.007

CellPress

Neuron
Article



# Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception

James O'Sullivan,[1] Jose Herrero,[3] Elliot Smith,[2,4] Catherine Schevon,[2] Guy M. McKhann,[2] Sameer A. Sheth,[2,5] Ashesh D. Mehta,[3] and Nima Mesgarani[1,6,*]
[1]Department of Electrical Engineering, Columbia University, New York, NY, USA
[2]Department of Neurological Surgery, The Neurological Institute, New York, NY, USA
[3]Department of Neurosurgery, Hofstra-Northwell School of Medicine and Feinstein Institute for Medical Research, Manhasset, New York, NY, USA
[4]Department of Neurosurgery, University of Utah, Salt Lake City, UT, USA
[5]Department of Neurosurgery, Baylor College of Medicine, Houston, TX, USA
[6]Lead Contact
*Correspondence: nima@ee.columbia.edu
https://doi.org/10.1016/j.neuron.2019.09.007

## SUMMARY

**Humans can easily focus on one speaker in a multi-talker acoustic environment, but how different areas of the human auditory cortex (AC) represent the acoustic components of mixed speech is unknown. We obtained invasive recordings from the primary and nonprimary AC in neurosurgical patients as they listened to multi-talker speech. We found that neural sites in the primary AC responded to individual speakers in the mixture and were relatively unchanged by attention. In contrast, neural sites in the nonprimary AC were less discerning of individual speakers but selectively represented the attended speaker. Moreover, the encoding of the attended speaker in the nonprimary AC was invariant to the degree of acoustic overlap with the unattended speaker. Finally, this emergent representation of attended speech in the nonprimary AC was linearly predictable from the primary AC responses. Our results reveal the neural computations underlying the hierarchical formation of auditory objects in human AC during multi-talker speech perception.**

## INTRODUCTION

In multi-talker acoustic environments, humans can easily focus their attention on one speaker even in the absence of any spatial separation between the talkers (Cherry, 1953). In such scenarios, the spectrotemporal acoustic components of the speakers are highly mixed at a listener's auditory periphery (Brungart et al., 2001). Successful perception of a particular speaker in this condition requires identifying and separating the spectrotemporal features of that speaker from the background and regrouping the acoustic components into a coherent auditory object that is unaffected by the variable acoustic overlap with other speakers (Bizley and Cohen, 2013; Shinn-Cunningham, 2008). The required neural computations that underlie this cognitive task in the human auditory system remain speculative, and this task has proved extremely challenging to model algorithmically (Luo and Mesgarani, 2019; Luo et al., 2018).

Studies on sound encoding in the mammalian auditory pathway have postulated the existence of a hierarchical, feedforward processing framework that starts from the auditory nerve and continues to primary and nonprimary auditory cortex (Hickok and Poeppel, 2007; Rauschecker, 1997). Neurons in this ascending auditory pathway have increasingly complex and multi-featured tuning properties (King and Nelken, 2009; Miller et al., 2002; Santoro et al., 2014). This encoding hierarchy results in a multidimensional and multiplexed representation of stimulus features in primary auditory areas that can facilitate auditory scene analysis (Patel et al., 2018; Walker et al., 2011). In parallel, substantial evidence indicates the existence of descending connections throughout the entire auditory pathway (Rasmussen, 1964). These task-dependent feedback connections modulate the tuning properties of auditory neurons, which likely support the extraction of target sound sources from the background acoustic scene (Fritz et al., 2003; Kilian-Hütten et al., 2011; Mesgarani et al., 2009a). The interaction between bottom-up and top-down mechanisms is particularly critical when attending to a target speaker in multi-talker acoustic conditions as the target and interfering sound sources overlap substantially in both time and frequency. Previous studies on multi-talker speech perception in the human auditory cortex have confirmed the emergence of a selective and enhanced representation of attended speech in higher auditory areas, such as superior temporal gyrus (STG; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013). Additionally, noninvasive studies have shown that attended and unattended talkers are co-represented in early components of neural responses, with distinct responses to the attended speaker appearing in only late response components and in only nonprimary auditory areas (Ding and Simon, 2012; Kerlin et al., 2010; Petkov et al., 2004; Power et al., 2012; Puvvada and Simon, 2017).

Although these findings suggest a progressive and hierarchical emergence of target speech from the mixed sound, how the primary and nonprimary auditory cortical areas represent mixed speech and how they interact to selectively enhance the target speech relative to the acoustic background remains unknown. In addition, whether these auditory cortical areas represent an attended speaker as an auditory object remains unclear

(Bizley and Cohen, 2013; Shinn-Cunningham, 2008). An auditory object representation implies invariance to the degree of spectrotemporal overlap with unattended speech, similar to the response permanence to partially occluded visual objects (Gibson, 2014). Although previous studies have shown a selective representation of attended speech in STG (Mesgarani and Chang, 2012), the difference between the neural responses to masked and unmasked spectrotemporal features of the attended speaker in primary and nonprimary areas is unknown.

To shed light on the encoding of mixed speech in primary and nonprimary auditory areas, we recorded from invasive electrodes implanted in patients undergoing neurosurgery as they focused on specific speakers in a multi-talker speech perception task. We used a combination of depth (stereotactic electroencephalogram [EEG]) and surface (subdural electrocorticography) recording techniques to reach both STG and Heschl's gyrus (HG). These speech-responsive areas (Khalighinejad et al., 2019; Mesgarani et al., 2014; Steinschneider et al., 2013) are easily identifiable from the macroscopic landmarks and are consistently present in all subjects, as opposed to the functional organization of auditory cortical fields which remains debated (Hackett et al., 2001; Moerel et al., 2014) and has a large intrasubject variability (Rademacher et al., 1993). While these regions are heterogeneous and each contain multiple auditory fields (Hamilton et al., 2018; Nourski, 2017), HG includes mostly the primary auditory cortex, and STG is considered mostly a nonprimary auditory area (Clarke and Morosan, 2012). Therefore, measuring the neural activity from both HG and STG areas allowed us to determine the encoding properties and functional relationship between these regions. Our results revealed significant differences between the representation of multi-talker speech in these two areas, a finding that contributes to a more complete functional and anatomical understanding of speech processing and auditory object formation in the human auditory cortex.

## RESULTS

Eight subjects participated in this study, with varying amounts of electrode coverage over their left and right auditory cortices. Two subjects had high-density electroencephalography (ECoG) grids implanted over their left temporal lobe with coverage of STG, and one of these subjects also had a stereotactic EEG (sEEG) depth electrode implanted in left HG. Depth electrodes with coverage over the left and right HG, as well as other auditory cortical regions, were implanted in the remaining 6 subjects. Figure 1A shows the electrodes from all subjects displayed on an average brain along with their corresponding measure of *effect size* (Cohen's D; Cohen, 2013) resulting from the comparison of the responses to speech versus silence (STAR Methods). Out of 624 electrodes, 230 were responsive to speech (effect size greater than 0.2), with 67 and 56 of these electrodes located in HG and STG, respectively. Further analyses were restricted to these electrodes in HG and STG (all electrodes are shown in Figure S1A).

### Stimuli and Example Responses

The subjects listened to stories read by a male speaker and female speaker, hereafter referred to as Spk1 and Spk2, respectively. The stimuli were presented in isolation (single-talker) and mixed together (multi-talker) with no spatial separation between them. The multi-talker condition was split into 4 blocks, and the subjects were instructed to pay attention to either Spk1 or Spk2 at the beginning of each block. The stories were intermittently paused, and the subjects were asked to repeat the last sentence of Spk1/2 in the single-talker condition or to repeat the last sentence of the attended speaker in the multi-talker condition to ensure that the subjects were engaged in the task. The performance for all subjects in the multi-talker condition was high (mean = 90%, STD = 8%, minimum = 80%).

Figure 1B shows portions of the stimuli and corresponding neural responses from 2 example electrodes in 1 subject, with one in STG (E1) and the other in HG (E2). By "response" here and in the rest of the manuscript, we are referring to the envelope of the high-gamma band (70–150 Hz; STAR Methods). The left panel shows a stimulus from the multi-talker condition (displayed as the superposition of the 2 speakers for visualization purposes). Qualitatively, the response of the neural site in STG (E1) changes depending on who is being attended, even though the stimulus is identical in both cases. Comparing the multi-talker responses with those obtained in the single-talker condition (middle and right panels) shows that the response of this site to the attended speaker is similar to the response to that speaker in isolation. Conversely, the neural site in HG (E2) responds similarly to the sound mixture irrespective of whether the subject is attending to Spk1 or Spk2. Comparing these responses with those in the single-talker condition suggests that the response of this site is the same as the response to Spk1 alone even when attending to Spk2. This visualization demonstrates the following response types: (1) sites that are modulated by attention to represent the attended speaker, and (2) sites that preferentially respond to a specific speaker even when not attending to that speaker. Motivated by this observation, we examined the extent to which each site was modulated by attention or was more responsive to one of the speakers.

### Selective Responses of Neural Sites to Specific Speakers

To study the preferential response of sites to the speakers across HG and STG, we examined the responses to the speakers in the single-talker condition. To compare the responses, we calculated the distribution of the normalized magnitude of the electrodes' response to Spk1 and Spk2. Figure 2A shows the response histograms for two example sites in HG. The difference between the medians of the distributions in Figure 2A confirms that these sites respond more strongly to Spk1 (left) or Spk2 (right). We quantified the preference for either speaker (the degree of the difference between the response distributions) by calculating the *effect size* (Cohen's D) of the difference. We term this metric the speaker-selectivity index (SSI); positive and negative values indicate a preference for Spk1 and Spk2, respectively. Evaluating the SSI (absolute values) across all electrodes revealed significantly more speaker-selective neural sites in HG than in STG (Figure 2B; unpaired t test, $p < 0.001$). This difference is also demonstrated by the wider distribution of the SSI in HG (SD of the SSI in HG and STG = 0.2 and 0.07, respectively). Figure S1B shows the spatial distribution of the SSI across the brain.

To examine the extent to which the observed preferred response to one speaker over the other can be explained by

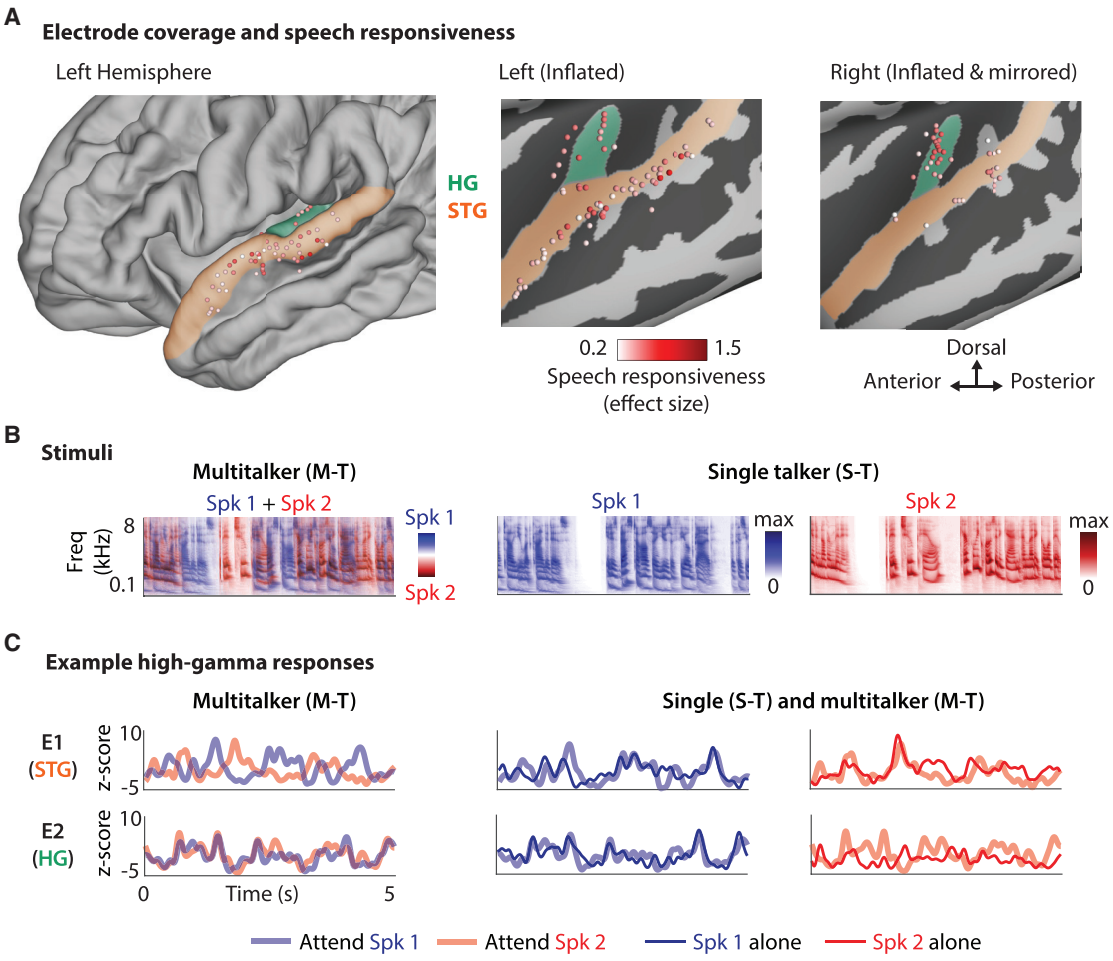


**Figure 1. Example of Neural Responses in Single- and Multi-talker Conditions**

(A) Electrode coverage and speech responsiveness. Electrodes from all 8 subjects were transformed onto an average brain. The left panel shows the left hemisphere, with HG (containing primary auditory cortex) highlighted in green, and STG (nonprimary auditory cortex) highlighted in orange. Middle and right panels show the inflated left and right hemispheres to assist visualization. The color of each electrode corresponds to the effect size (Cohen's D), measuring its response to speech versus silence. Only electrodes with an effect size >0.2 are shown.

(B) Stimuli. Portions of the stimuli (spectrograms) in the multi-talker (left) and single-talker (middle and right) panels. In the multi-talker condition, the spectrograms of Spk1 (male) and Spk2 (female) are superimposed for visualization purposes.

(C) Example neural responses from 2 electrodes in 1 subject: one in STG (e1) and the other in HG (e2). The response of e1 changes depending on which speaker is being attended, resembling the response to that speaker in isolation. Conversely, e2 responds similarly when attending to Spk1 and Spk2, as if it was responding to Spk1 alone, even when Spk2 is attended. This visualization demonstrates two response types: (1) sites with a modulated response to represent the attended speaker, and (2) sites that preferentially respond to one speaker irrespective of attention.

spectrotemporal tuning properties (Steinschneider et al., 2014), we first calculated the spectrotemporal receptive field (STRF) of each neural site. A STRF is a linear mapping between a stimulus (spectrogram) and the evoked response to that stimulus (Theunissen et al., 2000) that estimates the spectrotemporal features to which a neural site is tuned. The STRFs were calculated from the responses to the single-talker stimuli. Figure 2C displays the average STRFs from all electrodes that had an SSI either greater than +0.2 or less than −0.2 (selective for Spk1 or Spk2, respectively). To relate these tuning properties to the acoustic features of the speakers, we calculated the average acoustic spectrum of each speaker (labeled Spk1 and Spk2 Acous.; STAR Methods). To directly compare the acoustic spectrum of the speakers with each site's frequency tuning, we removed the temporal component of the STRFs by obtaining their 1$^{st}$ principal component (PC) along the spectral dimension. Therefore, we will abbreviate STRFs as spectral receptive fields (SRFs). The correlation (Pearson's r) between the SRFs that are selective for Spk1/2 and the spectral profile of Spk1/2 are 0.72 and 0.67, respectively ($p < 0.001$ for both; Figure 2D). The correlation between the difference in the SRFs and the difference in the spectral profile of the speakers is 0.82 ($p < 0.001$). These large correlation values suggest that the observed speaker selectivity of a neural site is largely due to a match between the spectral profile of the speakers and the frequency tuning of that site.

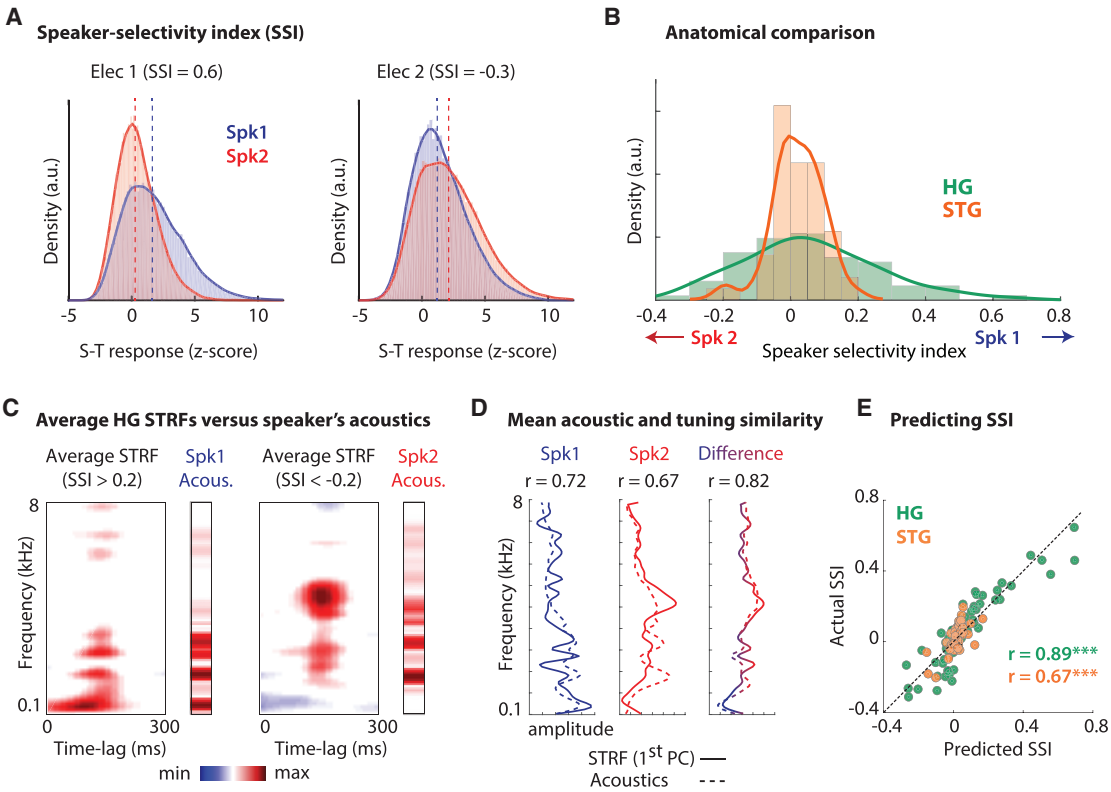


**Figure 2. Selective Responses of Neural Sites to Specific Speakers**

(A) The distribution of the responses to Spk1 and Spk2 in the single-talker condition from 2 example electrodes in HG. Electrodes 1 and 2 respond preferentially to Spk1 and Spk2, respectively. The dashed lines indicate the median of each distribution. The speaker selectivity index (SSI) is the effect size (Cohen's D) of the difference in the response to the 2 speakers. Positive numbers indicate a preference for Spk1, and vice versa.

(B) The distribution of the SSI in HG (green) and STG (orange) shows significantly more speaker-selective sites in HG ($p < 0.001$).

(C) Comparing the spectrotemporal tuning properties of neural sites with the acoustic profile of each speaker. Left panel: the average spectrotemporal receptive field (STRF) for all sites showing a preference for Spk1 (SSI >0.2) and the average acoustic spectrum of Spk1 (labeled Spk1 Acous.). Right panel: the average STRF for all sites showing a preference for Spk2 and the average acoustic spectrum of Spk2.

(D) The correlation between the average STRFs and average acoustics (after removing the temporal component of the STRFs by obtaining their 1st PC). Left panel: the correlation between the STRFs of Spk1 selective (SSI >0.2) sites (solid line) and the average acoustic spectrum of Spk1 (dashed line). Middle panel: the correlation between the STRFs of Spk2 selective sites and the average acoustics of Spk2. Right panel: the correlation between the difference in the 2 groups of STRFs and the difference in the acoustics of the 2 speakers.

(E) Predicting the SSI of a site from its STRF for all sites in HG (green) and STG (orange).

To examine the extent to which the SSI of each site could be predicted from its SRF, we used linear regression to map all sites in HG and STG from the SRF to SSI (STAR Methods; Figure S2). Figure 2E shows that speaker selectivity can be predicted for HG and STG electrodes with an accuracy of 0.89 and 0.67, respectively ($p < 0.001$ for both; Pearson's r value). The higher speaker preference prediction in HG indicates a more acoustically organized representation of the speakers in this area than that in STG. Together, these results suggest that sites in HG have more diverse spectral tuning properties, which results in an explicit representation of the distinct acoustic features of the two speakers.

## Attentional Modulation of Neural Responses

We showed that cortical areas have varied preferences for particular speakers and that sites in HG are more speaker selective than sites in STG. To determine the degree of attentional modulation of these sites, we compared the multi-talker and single-talker responses to measure how much the neural response to the mixed speech changed to resemble the response to the attended speaker in the single-talker condition (see Figure 1B, electrode e1 for an example). Therefore, we define the attentional modulation index (AMI) of neural sites as follows:

$$AMI = corr(Spk1_{attend}, Spk1_{alone}) - corr(Spk1_{attend}, Spk2_{alone}) + corr(Spk2_{attend}, Spk2_{alone}) - corr(Spk2_{attend}, Spk1_{alone})$$

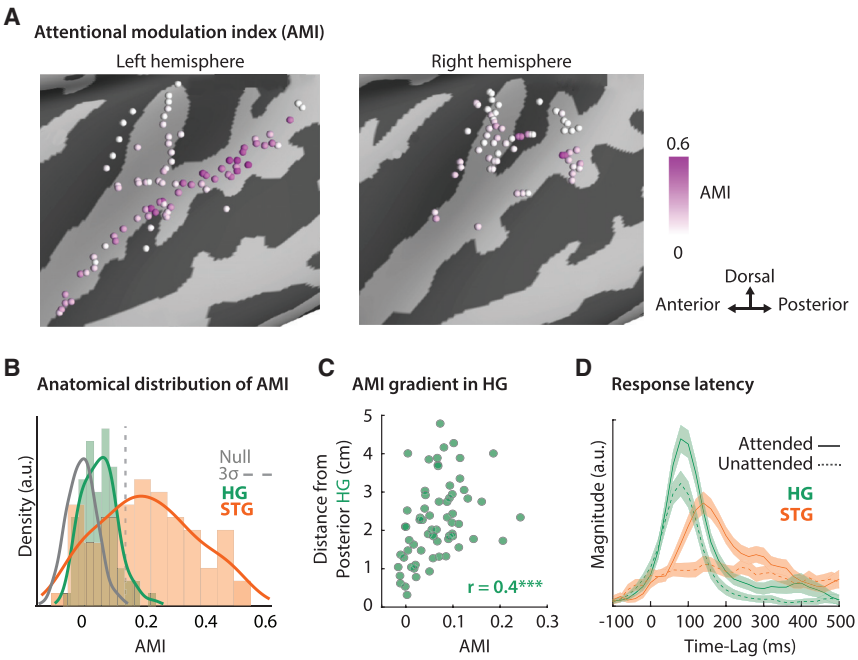


**Figure 3. Attentional Modulation of Neural Sites**

(A) The anatomical distribution of the AMI.

(B) The distribution of AMI in HG (green) and STG (orange) compared with a null distribution of the AMI (gray line). A significant AMI was defined as 3 times the standard deviation of the null distribution (3σ). Significantly more sites in STG (60%) than in HG (0.06%) are modulated by attention.

(C) The AMI of each site in HG compared with its distance from posterior HG. The positive correlation ($r = 0.4$, $p < 0.001$) demonstrates a gradient of attentional modulation emanating from this area.

(D) The response latency of the responses in HG (green) and STG (orange; mean ± SE) with respect to the attended (solid) and unattended (dashed) speakers. These plots were obtained by averaging the STRFs across frequency to obtain the temporal response profile for each site. This result demonstrates that STG sites respond later than do HG sites and shows greater suppression of the unattended speaker.

where *SpkX* refers to the response to speaker *X* either in the single-talker condition (alone) or when they are attended in the multi-talker condition (attend).

Larger AMI values indicate more attentional modulation of a neural site. Figure 3A displays the AMI across all neural sites that showed a significant response to speech. Figure 3B compares the AMI between HG, STG, and a null distribution obtained by randomly shuffling the trial order (gray line; STAR Methods). Figure 3B shows that a higher number of sites in STG are modulated by attention than those in HG, with 34 out of 56 sites in STG possessing an AMI significantly above chance (3σ; STAR Methods) compared with only 4 out of 67 sites in HG. This result shows that the representation in STG is more dynamic than that in HG and that the attentional state of the listener changes the representation in STG more than in HG. Notably, the population of HG sites as a whole had a distribution of AMI significantly above that of the null distribution (unpaired t test; $p < 0.001$), suggesting a weak but significant effect of attention in HG. In addition, Figure 3C shows a linear increase in the AMI with increasing distance from posterior HG (MNI coordinates: $x = 35$, $y = -30$ and $z = 18$; $r = 0.4$, $p < 0.001$). This finding reveals a gradient of attentional modulation from posterior HG toward STG. Figure 3D shows the latency of the responses in HG and STG (mean ± SE) with respect to the attended (solid) and unattended (dashed) speakers. These response latencies were obtained by averaging the STRFs across frequency to obtain the temporal response profile for each site. This finding shows that STG responds later than HG and further illustrates the greater suppression of the unattended speaker in STG.

To examine the relationship between the speaker selectivity of sites and their degree of attentional modulation, we calculated the joint distribution of the SSI and AMI (Figure 4A), comparing STG (orange) and HG (green). Figure 4B displays the AMI and SSI across all neural sites that showed a significant response to speech. These plots illustrate a fundamental difference between the organization of the neural responses in HG and STG where HG is relatively static and responds preferentially to speaker differences, whereas STG favorably represents the attended speaker.

## Emergence of Auditory Objects: The Neural Representation of Masked versus Unmasked Acoustic Features

Motivated by the clear difference between the organization of responses to multi-talker speech in HG and STG (Figure 4), we further examined the similarity of the neural responses to speakers from the single to multi-talker conditions. The speech signal varies across both time and frequency; therefore, the spectrotemporal features of an attended speaker variably overlaps with interfering speakers (Figure 1B). Here, we examined how the variable overlap between attended and unattended speakers affected the neural responses in HG and STG.

The overlap between two competing speakers is easy to quantify in the time-frequency domain (i.e., the spectrogram); however, the neurons in auditory cortex can have complex and often nonlinear tuning properties, making it difficult to assess the degree of overlap between the features to which they are tuned. To circumvent this problem, we developed a model-independent method to evaluate neural responses as a function of the relative energy of both speakers with respect to the feature to which a neural site is tuned. To accomplish this goal, we superimposed the magnitude of the responses in the multi-talker condition onto the joint distribution of the responses to Spk1 and Spk2 alone. Figure 5A shows an example STG electrode. The top panel shows the responses in the S-T condition to Spk1 (blue) and Spk2 (red). The bottom panel shows the responses in the M-T condition when Spk1 is attended (top) or when Spk2 is

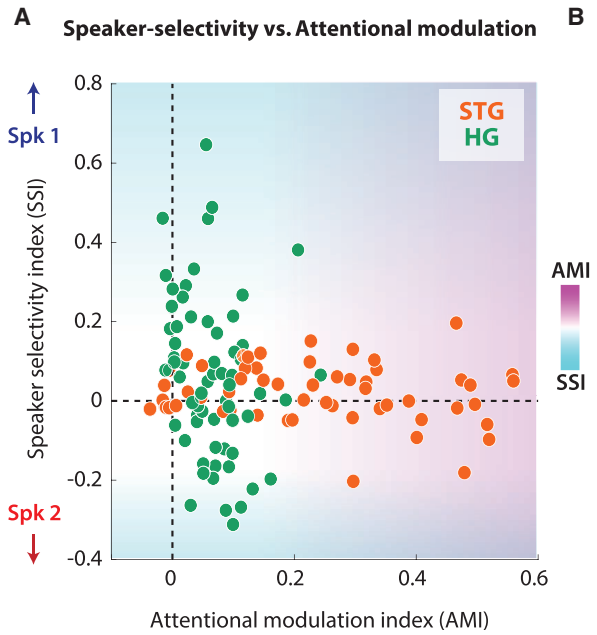


**Figure 4. Speaker-Selectivity Index versus Attention-Modulation Index (AMI)**

(A) The joint distribution of the AMI (x axis) and SSI (y axis) in HG (green) and STG (orange). This distribution further illustrates that HG shows the small effect of attention and a large amount of speaker selectivity. Conversely, STG exhibits a large effect of attention and little speaker selectivity.

(B) The anatomical distribution of the SSI (cyan) and AMI (magenta). These plots illustrate a fundamental difference between the nature of the representation in HG and STG where HG provides a feature-rich, relatively static representation of the speakers, whereas STG filters out the unwanted source and selectively represents the attended speaker.

attended (bottom). The color in these cases represents the amplitude of the response. Three time points are denoted (a, b, and c). The top-right panel shows the 2D histogram of the joint distribution of the responses to Spk1 (x axis) and Spk2 (y axis) in the S-T condition. The 3 time points (a, b, and c) are marked. In the bottom-right panel, the response amplitude of the M-T condition is superimposed on the S-T histogram (from above). The color corresponds to the response amplitude in the M-T condition. This calculation is performed separately for each attention condition (A1: attend Spk1, and A2: attend Spk2). For this example STG electrode, the representation rotates 90 degrees when the attended speaker changes. This change can be summarized by adding A1 to the transpose of A2 to obtain a single matrix that shows the magnitude of the multi-talker response with respect to the attended and unattended speakers (Figure 5B). The rows of this matrix show the response to the attended speaker as the magnitude of the unattended speaker varies (changing colors). This finding reveals that this site responds as a linear function of the attended speaker. However, there is an effect of energetic masking at the extrema when the magnitude of the unattended speaker is very large relative to the attended speaker (dark lines, left of figure). Alternatively, the columns of the matrix show that this site is almost unaffected by the magnitude of the unattended speaker except when the magnitude of the attended speaker is very small (light colored lines). To summarize the response of this electrode, we took the average across the rows and columns (Figure 5B, next-to-last right panel). In summary, this site responds as a linear function of the features of the attended speaker, meaning that louder features of the attended speaker result in a larger response. At the same time, this site is mostly unaffected by the unattended speaker, meaning that despite the change in the overlap between the features of the unattended and attended speakers, this change in masking is not reflected in the responses. Figure 5B (bottom panels) shows the same analyses of an example electrode in HG. This neural site appears to be unaffected by attention, responding linearly with respect to both speakers. This observation means that this site responds to the acoustic feature to which it is tuned irrespective of whether that feature belongs to the attended or unattended speaker. However, we observed a slight effect of attention at the extrema, which is further illustrated in Figure S6.

Figure 5B (right-most panel) illustrates the effect of masking across the population of neural sites in HG and STG. This analysis reveals that (1) STG sites respond to the acoustic features of the attended speaker and are unaffected by how much these features are masked by the unattended speaker. (2) HG sites respond to the features of both speakers. Although previous studies have postulated that attention may act as a linear gain change to enhance attended and suppress unattended speakers, our findings show that the unattended speaker is not simply attenuated (which would result in a linear interaction with the attended speaker) but is nonlinearly suppressed in STG responses. This nonlinear effect is quantified in Figure S6C where we calculate a linear fit to each masking curve in HG and STG (i.e., Figure 5). A linear fit performs well for attended speech in HG (goodness of fit [GOF] = median ± STD: 0.98 ± 0.07) and STG (GOF = 0.98 ± 0.14) and unattended speech in HG (GOF = 0.96 ± 0.12). However, a linear fit performs poorly for unattended speech in STG (GOF = 0.63 ± 0.28). The linear response of HG sites to the degree of masking indicates the acoustic nature of the representation in this region with no evidence for feature grouping. Nonetheless, the nonlinear suppression of masking in STG responses indicates that the speaker features are grouped and represented as a coherent auditory object in this area.

## Separability of Speakers in the Population Activity of HG

We have demonstrated an acoustic and linear representation of mixed speakers in HG with relatively small attentional modulation effects. Primary auditory cortex, however, is several synapses away from the auditory periphery. Hence, the speech signal must go through a series of transformations before it gets to the primary auditory cortex (Webster and Fay, 2013). To shed light on the encoding properties of the population responses in HG

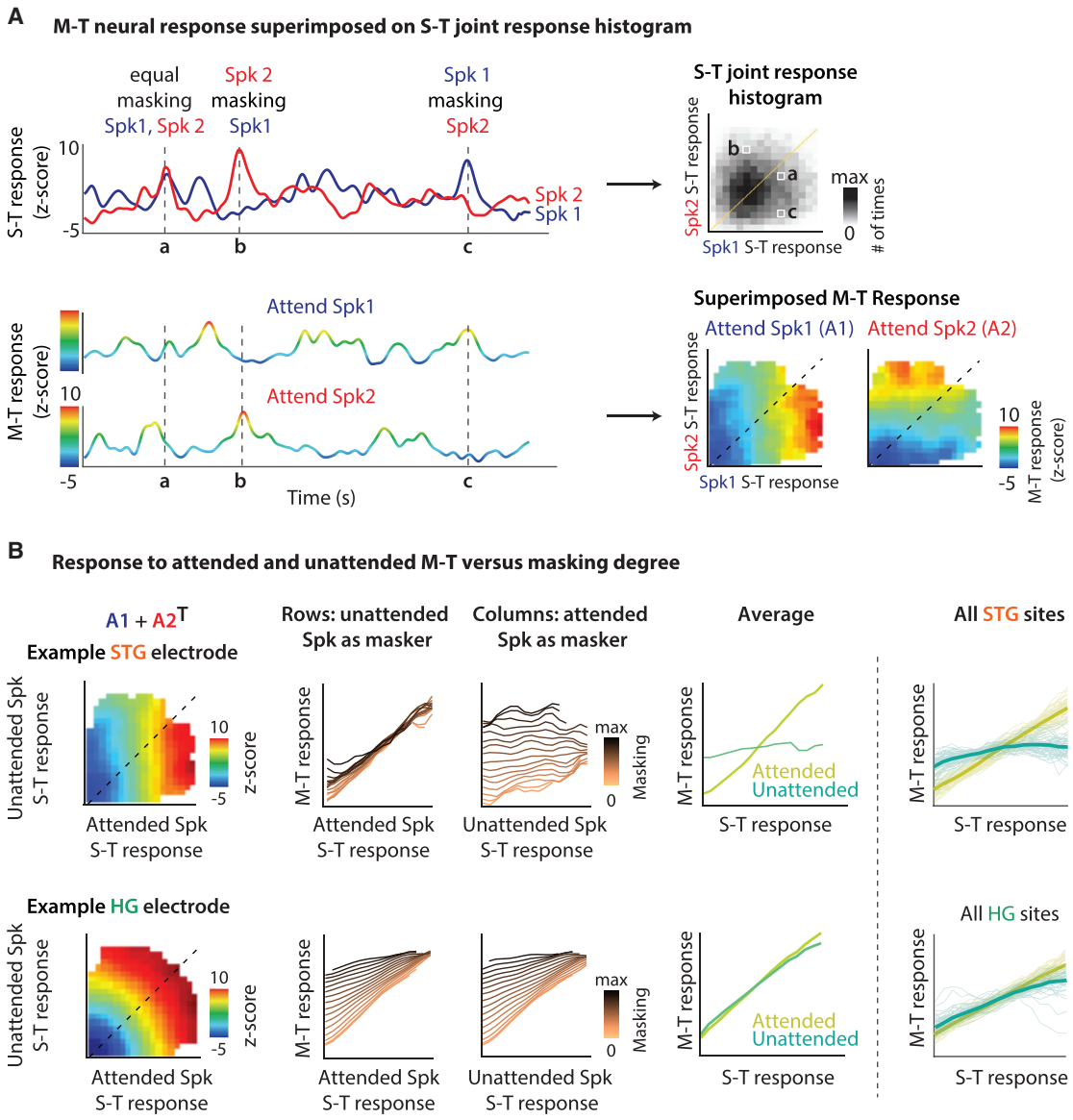


**Figure 5. The Representation of Auditory Objects in HG and STG**

The magnitude of the responses in the multi-talker (M-T) condition are superimposed onto the joint distribution of the responses to Spk1 and Spk2 in the single-talker (S-T) condition.

(A) For an example STG electrode, the top panel shows the responses in the S-T condition to Spk1 (blue) and Spk2 (red). The bottom panel shows the responses in the M-T condition when Spk1 is attended (top) or when Spk2 is attended (bottom). The color in these cases represents the magnitude of the response. Three time points are denoted (a, b, and c). The top-right panel shows the 2D histogram of the joint distribution of the responses to Spk1 (x axis) and Spk2 (y axis) in the S-T condition. The 3 time points (a, b, and c) and marked. In the bottom-right panel, the response magnitude of the M-T condition is superimposed on the S-T histogram (from above). The color corresponds to the response magnitude in the M-T condition. This calculation is performed separately for each attention condition (A1: attend Spk1, and A2: attend Spk2), illustrating a large effect of attention as the representation rotates 90 degrees.

(B) Summarizing the responses by adding A1 to the transpose of A2. The rows of this matrix show the response to the attended speaker as the magnitude of the unattended speaker varies (changing colors), and the columns show the response to the unattended speaker as the magnitude of the attended speaker varies. This finding reveals that this site responds as a linear function of the attended speaker and is almost unaffected by the magnitude of the unattended speaker. Taking the average across the rows and columns allows for a summary of this response type (right panel). The bottom panels show the same analysis for an example electrode in HG. This finding reveals that this neural site appears to be unaffected by attention, responding linearly with respect to both speakers. The right-most panels show the average summary plots across the population of neural sites in HG and STG. This analysis reveals that (1) STG sites respond to the acoustic features of the attended speaker and are unaffected by how much these features overlap by the unattended speaker, providing evidence for the grouping of features of the attended speaker. (2) HG sites respond to the features of both speakers with no evidence of a coherent response to attended speaker features.

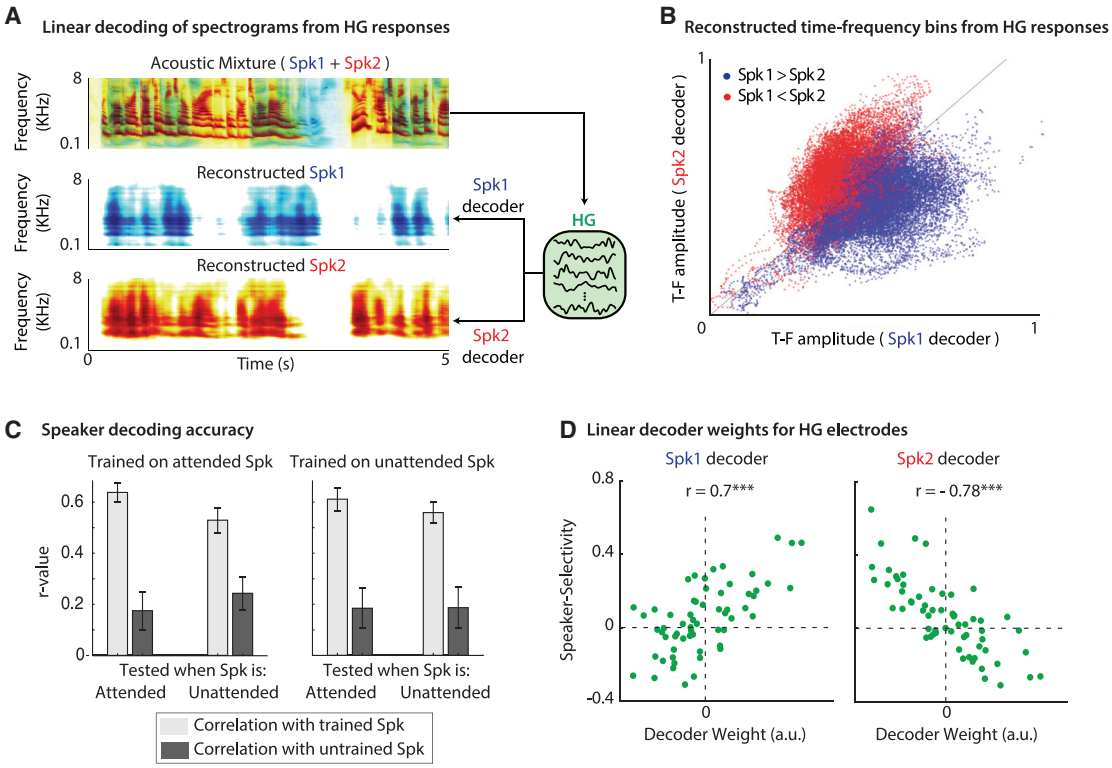


**Figure 6. Speakers Are Linearly Separable in HG**

(A) Training linear decoders to extract either speaker from the representation of the mixture in HG. Top panel: the spectrogram of the mixture (displayed as the superposition of Spk1 and Spk2). Linear decoders can reconstruct either Spk1 (middle) or Spk2 (bottom) from the neural responses in HG to the mixture.

(B) Scatterplot of the amplitude of all time-frequency (TF) bins when reconstructing Spk1 (x axis) versus reconstructing Spk2 (y axis). The dots are colored according to the dominant speaker in the corresponding T-F bin.

(C) Irrespective of the actual attended speaker, both speakers can be extracted from the representation of the mixture in HG. Left panel: decoders were trained on the attended speaker and tested when that speaker was either attended or ignored (see x labels). Right panel: decoders were trained on the ignored (unattended) speaker and tested when that speaker was either attended or ignored (see x labels). Light gray bars indicate the correlation (mean ± STD) with the trained speaker, and dark gray bars indicate the correlation with the untrained speaker. In all cases, the reconstruction has a significantly higher correlation ($p < 0.001$) with the trained speaker than with the untrained speaker.

(D) The SSI for each electrode in HG (green dots) is plotted against the average weight that the decoders learn to apply to them when the decoders are tasked with extracting Spk1 (left panel) or Spk2 (right panel). The decoders learn to enhance/suppress the electrodes that are selective for Spk1/Spk2 depending on the speaker to be extracted.

to mixed speech, we tested how the population of neural responses in HG can support speaker separation. We used a rudimentary linear decoder to measure how well the clean speaker spectrograms can be extracted from the HG responses to the mixed speech. To do so, we used a method known as stimulus reconstruction, which finds a linear mapping (decoder) between a stimulus (spectrogram) and corresponding evoked neural responses (Akbari et al., 2019; Mesgarani et al., 2009b).

We used the stimulus reconstruction method to decode the representation of Spk1 and Spk2 from HG responses to the mixture, as shown in Figure 6A. The top figure shows the acoustic mixture (for visualization purposes, Spk1 [blue] is superimposed on Spk2 [red]). The middle and bottom panels show example reconstructed spectrograms from the neural responses to mixed speech in HG when the decoders were trained to map the neural responses to the clean spectrograms of Spk1 (middle panel) or Spk2 (bottom panel; Figure S4A). The high correlation between the actual and reconstructed spectrograms for Spk1 (0.64, $p < 0.001$) and Spk2 (0.65, $p < 0.001$) shows that the speakers are highly separable in the population activity of HG.

Although the high reconstruction accuracy shows faithful decoding of the spectrotemporal features of each speaker, it does not specify which time-frequency components are more decodable. Figure 6B shows a scatterplot of the reconstructed spectrograms of the same mixture sound from the reconstruction decoder trained on Spk1 (x axis) plotted against the reconstruction decoder trained on Spk2 (y axis). Each dot represents a time-frequency bin of the reconstructed spectrogram and is colored according to the relative magnitude of the speakers in that bin in the original acoustic mixture. Blue dots correspond to a time-frequency bin in which the magnitude of Spk1 was greater than that of Spk2 and vice versa (STAR Methods; Figure S4B). The separability of red and blue dots in Figure 6B shows that the linear model can correctly pull out the time-frequency bins of Spk1 and Spk2. This observation further supports the notion that HG responses give rise to a representation of the

mixture in which the acoustic features of each speaker become readily decodable.

The previous analyses showed a diverse and explicit representation of the speakers at individual HG sites and that the speakers are separable at the population level. To relate the local and population encoding properties in HG, we examined whether the neural sites that are highly tuned to the acoustic features of each speaker are responsible for successfully reconstructing the individual speakers from the mixture. Since a linear decoder is a spatiotemporal filter that applies a weight to each electrode at a specified number of time lags, we can gain insights into how a decoder learns to separate the speakers by examining these weights. Figure 6D displays the weight applied to each electrode plotted against the SSI for that electrode (to obtain a single weight for each electrode, we averaged the weights across frequency and time, as well as across attention conditions). As shown, the decoders learn to place larger weights on the speaker-specific electrodes and alternate the weights depending on the speaker to be extracted ($r = 0.7$ and $r = -0.78$, when trained to extract Spk1 and Spk2, respectively). This result shows the high contribution of speaker-selective sites in HG to decoding the individual speakers from the responses to the mixture.

The successful decoding of speakers from the HG responses to the mixture suggests that the representation of mixed speech in HG may serve as a basis for higher auditory areas, such as STG, in which the attended speaker can be extracted by changing the weights from the HG responses. However, for this computation to work, the readout of a specific speaker from specific HG sites should not be affected by the attentional state. Otherwise, the decoding scheme would also need to be updated as the listener switches attention. To examine whether speaker decoding accuracy depends on the attentional state, we trained and tested linear decoders from all possible combinations of training/testing and attention (Figure 6C). The left panel illustrates the decoders that were trained on the attended speaker. The light gray bars display the correlation between the reconstructed and actual spectrogram of the speaker on which the decoder was trained. The dark gray bars display the correlation between the reconstructed and actual spectrogram of the speaker on which the decoder was not trained. The x axis is partitioned into instances when the trained speaker was either attended or ignored during testing. Each decoder was trained on the clean spectrogram of the attended speaker on a portion of the data (4-fold cross-validation). This decoder was then used to reconstruct a spectrogram from 2 different test sets when (1) the trained speaker was attended to and (2) the trained speaker was ignored. The right panel illustrates a similar combination of training and testing, and the only difference is that the decoders were trained on the ignored speakers. The small change in reconstruction accuracy as attention switches demonstrates that a decoder that is trained to separate a speaker from the mixture of responses in HG generalizes well to the condition where that speaker is attended to or ignored. By training a decoder on data when a speaker is attended but testing the decoder on data when that speaker is ignored, we have shown that the decoding scheme that is required to segregate a speaker from HG responses remains unchanged, which is an important property of the representation because it enables the constancy of decoding the sound sources from HG. Although we do see a small effect of attention when decoding a speaker from HG responses (the difference between the decoding of a speaker in the attended/ignored condition; $p < 0.001$, t test), this effect is likely caused by the small effect of attention on HG that we showed previously (Figure 3B).

## Emergent Representation of Attended Speech in STG

The exact connectivity between HG and STG is not yet fully established in humans, yet ample evidence suggests that HG includes the primary auditory cortex, whereas STG contains mostly non-primary belt and parabelt areas (Hackett, 2008; Moerel et al., 2014). Consistent with this notion, STG sites in our study had significantly longer response latencies than did HG sites (Figure 7A; STAR Methods). To estimate the response latency of an electrode, we squared its STRF, averaged across frequency, and measured the latency as the peak magnitude of the result. In addition, the responses in STG were significantly better predicted from the responses in HG than vice versa (Figure 7B; unpaired t test, $p = 0.033$; STAR Methods). This observation was made despite the significantly better prediction of responses by STRFs in HG than in STG (Figure 7B; unpaired t test, $p = 0.016$). These results suggest that the STG sites in our study may be more downstream relative to the HG sites, which is consistent with the architectonic studies of these regions (Clarke and Morosan, 2012).

To examine whether the representation of attended speech in STG can be predicted from the responses in HG, we used linear regression to estimate the responses in STG from the population of HG sites separately for when Spk1 or Spk2 was being attended. Figure 7C (left panel) shows the results of this analysis for an example electrode in STG. Each green dot on the left of the figure represents an electrode in HG, and the orange dot on the right represents one example electrode in STG. The electrodes are plotted according to their SSI and AMI (similar to the plot in Figure 4A). The color of the lines connecting HG electrodes to STG electrodes indicates the change in prediction weight between the 2 attention conditions (see Figure S5A for the weights in each attention case). As shown, the largest weight changes correspond to the most speaker-selective sites. The correlation between the weight change and SSI for this electrode is 0.69 (Pearson's r). Figure 7C (right panel) shows the change in weights for all STG sites, illustrating a consistent effect across the population. For all STG electrodes, the correlation between the average weight change and SSI is 0.83 ($p < 0.001$; Figure 7D). In addition, we found a strong correlation between the AMI of an STG site and the change in HG weights ($r = 0.54$, $p < 0.001$; Figure S5B). This dynamic modulation of the weights from HG suggests a possible computational mechanism for the selective representation of the attended speaker and the suppression of the unattended speaker in STG (Mesgarani and Chang, 2012). That is, STG sites may change their synaptic weight to increase the input from HG electrodes that are selective for the attended speaker and decrease the input from HG electrodes that are selective for the unattended speaker.

Our proposed computational model requires known decoding weights for STG sites from HG. Even though we showed that the decoding weights are highly correlated with the SSI of each site, the SSI of each site in the multi-talker condition is not given. To determine whether speaker decoding weights given to each

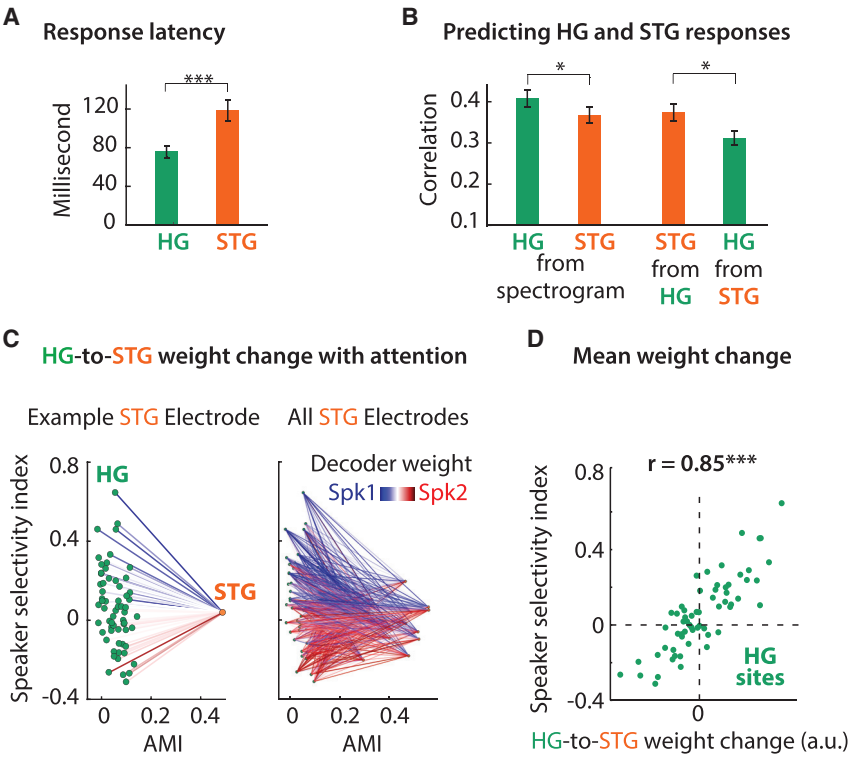


**Figure 7. Mapping HG to STG**

(A) STG (orange) responds with a longer latency than HG (green), suggesting that STG is further downstream (cf. Figure 3D).

(B) The neural responses in HG and STG can be predicted from the acoustic spectrogram (using a STRF) or from each other. Both areas can be predicted from the stimulus (left panel), with HG having significantly higher ($p < 0.05$) prediction accuracies. However, when mapping from HG to STG (and vice versa), HG can predict STG significantly better than STG can predict HG ($p < 0.05$). Error bars denote the mean ± SE. Data are from the single-talker condition.

(C) Mapping HG to STG in the multi-talker condition. Left panel: for an example electrode in STG (orange dot), under attention, the weights from each HG electrode (green dots) change to enhance (suppress) the attended (unattended) speaker. Blue (red) lines correspond to a larger weight when Spk1 (Spk2) is attended.

(D) The average weight change for each HG electrode (green dots) plotted against their corresponding SSI. The positive correlation ($r = 0.85$) confirms that larger weight changes are applied to the most speaker-selective sites in HG.

electrode in HG can be determined in an unsupervised manner, we tested a plausible mechanism known as the temporal-coherence model of stream segregation (Shamma et al., 2011). This theory posits that mixed sources can be segregated because the various constituent components of a single source will be correlated over time and uncorrelated with the components of another source. In the case of the two speakers, the neural sites in HG that are selective for Spk1 should be uncorrelated with those that are selective for Spk2. Figure 8A shows the correlation between all HG electrodes over time sorted according to their SSI (STAR Methods). Figure 8B shows the magnitude of the sum of the first three principle components (PCs) of this matrix, plotted against the corresponding SSI for each electrode. The high correlation ($r = 0.87$, $p < 0.001$) demonstrates that temporal coherence highly predicts the speaker selectivity of neural sites. In addition, the high correlation ($r = 0.79$, $p < 0.001$; Figure 8C) between these PCs and corresponding HG weight changes for speakers shows that the linear weights required to separate a particular speaker from the mixed HG representation can be found automatically, without needing any prior knowledge or supervision.

## DISCUSSION

By leveraging invasive neural recordings from the human auditory cortex, we examined the hierarchical and progressive extraction of attended speech in a multi-talker scenario. The high spatial and temporal resolution of our method allowed us to determine the encoding properties of target and interfering sources from primary (posterior HG) to nonprimary (STG) auditory cortical areas and to relate the representation of mixed speech between these regions that leads to the enhanced encoding of attended speech. Specifically, based on our findings, HG has more diverse spectrotemporal tuning properties than does STG, which results in more selectivity for the distinct features of individual speakers. The HG representation is also relatively static, i.e., showing little effect of attention. However, the population of responses in HG support a simple readout of the individual speakers in the mixture. In contrast, STG is more dynamic and selectively encodes the acoustic features of the attended speaker. Moreover, by examining the degree of acoustic overlap between the target and interfering speakers, we found that STG (but not HG) nonlinearly suppresses the overlapping features of interfering sources, which results in an invariant representation of the target speaker. Finally, we examined the relationship between the representation in HG and STG using a linear model and successfully accounted for the formation of the target speaker representation in STG from HG. In this model, attention changes the weights of the input from HG to STG to utilize the speaker-selective sites in HG to extract either speaker. Importantly, these weights can be determined solely from the temporal coherence of the neural activity in HG.

Our results show a stark contrast between the encoding properties of multi-talker speech in HG and STG where HG creates a rich representation of the mixed sound, and STG invariantly represents the attended source. We showed that representation of mixed speech in HG enables decoding of both attended and unattended speakers and may facilitate their extraction in downstream cortical areas (Puschmann et al., 2018). The neural transformations of the acoustic signal that enable such a representation in HG remain an open question. Previous research has

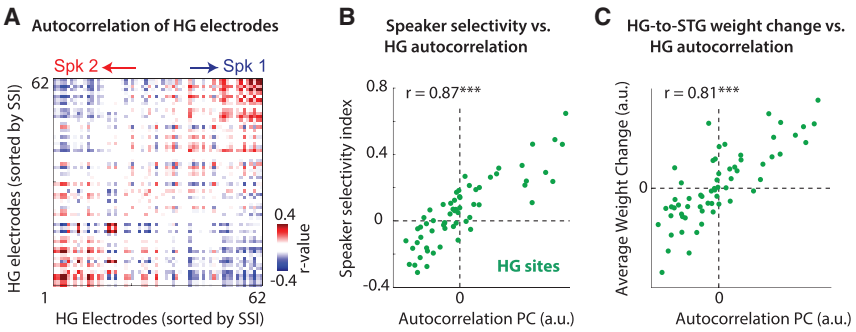


**Figure 8. Determining Speaker Selectivity in the Multi-talker Condition**

Given only the representation of the mixture in HG, sites that are selective for either speaker can be determined by obtaining the correlation structure (temporal coherence) of the responses.

(A) The correlation between all HG sites sorted according to their SSI.

(B) Decomposing the correlation matrix in (A) using principal-component analysis (PCA) permits the acquisition of a single number for each site. The large correlation (r = 0.87) with the corresponding SSI for each electrode demonstrates that the SSI can be obtained from the multi-talker responses alone.

(C) Similarly, the weights from HG to STG (HGRF) in the multi-talker condition can be determined from the same PCA analysis (r = 0.81; cf. Figure 6D).

shown a hierarchical transformation of the acoustic signal as it travels from the auditory nerve to primary and nonprimary auditory cortical areas (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). Specifically, neurons in the ascending auditory pathway are tuned to increasingly more complex and multi-featured spectrotemporal patterns (Linden et al., 2003; Miller et al., 2001, 2002), which results in a multidimensional and joint encoding of a multitude of acoustic dimensions in primary auditory areas (Bizley et al., 2009; Mesgarani et al., 2008; Patel et al., 2018; Walker et al., 2011). This increasingly complex tuning to multiple acoustic features results in an explicit representation of the spectrotemporal features of an acoustic stimulus from which the informative aspects of that stimulus can be more easily decoded (Han et al., 2019; Mesgarani et al., 2008; O'Sullivan et al., 2017; Walker et al., 2011). Consistent with our findings, the multidimensional representation of stimuli in early auditory areas supports the grouping of coherent acoustic dimensions and the formation of auditory objects in higher auditory areas where categorical and perception-driven representations of acoustic stimuli emerge (Bidelman et al., 2013; Bizley et al., 2013; Chang et al., 2010; Elhilali et al., 2009; Leaver and Rauschecker, 2010; Nourski et al., 2015, 2019; Teki et al., 2016). The stimulus cues used by the auditory pathway to enable this grouping are not completely clear. Computational models of speech separation (Luo and Mesgarani, 2019; Luo et al., 2018; Mesgarani et al., 2010) have shown the efficacy of several acoustic cues, including common onset and offset, spectral profile, and harmonicity. In addition, areas in STG encode linguistic cues (Mesgarani et al., 2014), which could be used to facilitate segregation or the recovery of masked features (Leonard et al., 2016), including phonotactic probability (Brodbeck et al., 2018; Leonard et al., 2015; Di Liberto et al., 2019), syntax (Fedorenko et al., 2016; Nelson et al., 2017), and semantics (Broderick et al., 2018; de Heer et al., 2017; Huth et al., 2016). Future studies that directly manipulate the linguistic structure of sentences in multi-talker conditions are needed to investigate the extent to which these cues may be used for speaker segregation in the human auditory cortex. In addition, what top-down mechanisms drive speaker segregation in STG remains an open question. The fronto-parietal attention network has shown to play a role, in particular, the frontal eye fields, the temporoparietal junction, and the intraparietal sulcus (Hill and Miller, 2010; Lee et al., 2014; Molenberghs et al., 2007; Salmi et al., 2009). Future studies with invasive electrodes in these areas may provide further information on their mechanistic contribution to speaker segregation.

We found a gradient of attentional modulation from posterior to anterior HG that continued toward the posterior STG. This finding is consistent with the anatomical and functional organization studies of the human auditory cortex, which suggests that primary auditory cortex originates in posterior HG, with belt/parabelt regions extending to anterior HG and STG (De Martino et al., 2015; Moerel et al., 2014; Morosan et al., 2001; Nourski, 2017; Steinschneider et al., 2014). Previous studies have found a similar organization of attentional modulation in humans (Nourski et al., 2017; Obleser et al., 2007; Petkov et al., 2004; Puvvada and Simon, 2017; Steinschneider et al., 2014). Our choice of focusing on the anatomical division of HG and STG, however, is functionally imprecise, because HG is not a single functional area and anterolateral HG may be higher in the cortical hierarchy than portions of STG (De Martino et al., 2015; Moerel et al., 2014; Nourski et al., 2017). In a similar vein, STG may also contain further subfields (Hamilton et al., 2018). Future studies with higher density neural recordings (Khodagholy et al., 2015) from these areas can further tease apart the response properties within each cortical region and provide information that is critically needed to fully describe the functional organization of human auditory cortex.

Similar to our finding of higher attentional modulation of responses in STG compared to HG, animal studies have also reported substantially more enhanced responses to target stimuli in secondary auditory areas compared to the primary auditory cortex (Atiani et al., 2014) and subcortical areas (Slee and David, 2015). These studies, however, reported a higher attentional effects in primary areas compared to what we observed in HG. One possible reason for this bigger effect could be the simplicity of the stimuli used in those studies (e.g., pure tones) where the target and interfering sounds were separable even along the tonotopic axis at the auditory periphery. The combined evidence suggests that the attentional modulation of the neural representation of target sound sources may occur only at a level of representation where the tuning properties of the neurons has enough capacity to realize sound separation. Additional experiments comparing the attentional modulation of various auditory cortical areas with tasks that systematically increase the

spectrotemporal overlap of the target and background scene can further shed light on this hypothesis. An alternative explanation is the difference between the behavioral demands and the ecological relevance of the stimulus used in our study compared to animal studies (Atiani et al., 2009), which may differently recruit the neural circuits of attention which varies as the task's reward structure changes (David et al., 2012). Since the performance of the subjects in our task was close to ceiling, we could not study the effect of behavioral performance on the modulation of neural responses. However, previous studies have found correlates of behavioral failure in the neural data recorded from the human STG (Mesgarani and Chang, 2012).

We demonstrated that a linear model can successfully map the responses in HG to those in STG, and this connection can account for the attentional modulation of STG responses from HG in the multi-talker condition. Importantly, the required changes in the weights of the model can be found simply from the temporal correlation of the neural activity in HG (Krishnan et al., 2014; O'Sullivan et al., 2015; Shamma et al., 2011; Thakur et al., 2015). Although debates regarding the anatomical and functional connectivity of these two regions are ongoing (Moerel et al., 2014), recent fMRI work supports a hierarchical model of speech processing progressing from HG to STG and beyond (de Heer et al., 2017). In addition, intracranial recordings have shown functional connectivity between HG and STG, with bottom-up information transfer observed in a similar frequency band as in our study (higher than 40 Hz; Fontolan et al., 2014). Future research that tests the causal relationship between these two regions, for example, by using cortico-cortical evoked potentials (CCEPs; Keller et al., 2014) during multi-talker speech perception, might be particularly suitable to shed light on the information transfer and dynamic connectivity of these areas as the attentional focus of the subject changes. We did not find any significant differences between the two hemispheres with regards to speaker selectivity or attentional modulation (Figure S7B), which could be due to the lack of enough anatomical coverage. Future studies with more extensive coverage may be able to shed light on potential hemispheric differences in the acoustic processing of speech (Flinker et al., 2019).

We tested the formation of auditory objects in HG and STG by examining the responses to target features in the presence of variable overlap with the interfering speaker (Bizley and Cohen, 2013; Shamma, 2008; Shinn-Cunningham, 2008). We found that HG sites respond to the total sum of spectrotemporal energy in the acoustic signal irrespective of whether the energy belongs to the target or interfering speaker. This observation confirms that HG does not represent segregated and grouped spectrotemporal features of target sound sources; hence, the attended auditory objects are not yet formed in this region. However, STG showed nonlinear suppression of the acoustic overlap with the interfering source, resulting in an invariant representation of the attended features that is unaffected by the amount of acoustic overlap, indicating the presence of auditory objects in this region. These findings are consistent with previous noninvasive studies that showed a late emergence of attended speech (Ding and Simon, 2012; Power et al., 2012) in only higher auditory regions (Petkov et al., 2004). However, the high temporal and spatial resolution of our recording method allowed us to further determine the encoding properties of target and interfering sound sources in these areas.

By examining the representational and encoding properties of speech in primary and nonprimary auditory areas, our study takes a major step toward determining the neural computations underlying multi-talker speech perception and the interaction between bottom-up and top-down signal transformations that occur in the auditory pathway that gives rise to a segregated and grouped representation of attended auditory objects.

## STAR★METHODS

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

- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human Subjects
  - Stimuli and Experiments
- METHOD DETAILS
  - Data Preprocessing and Hardware
  - Acoustic Spectrum of Speakers
  - STRFs and Stimulus Reconstruction
  - Predicting Speaker Selectivity and Attentional Modulation
  - Mapping HG to STG
  - Temporal Coherence
  - Transformation of Electrode Locations onto an Average Brain
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Speaker-Selectivity Index (SSI)
  - Attentional Modulation Index (AMI)
- DATA AND CODE AVAILABILITY

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.neuron.2019.09.007.

## ACKNOWLEDGMENTS

This work was funded by a grant from the NIH, NIDCD-DC014279, S10 OD018211, and the Pew Charitable Trusts, Pew Biomedical Scholars Program.

## AUTHOR CONTRIBUTIONS

J.O. and N.M. designed the experiment. J.O., J.H., E.S., C.S., G.M.M., S.A.S., A.D.M., and N.M. recorded the neural data. J.O. and N.M. analyzed the data and wrote the manuscript. All authors commented on the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 28, 2019
Revised: July 11, 2019
Accepted: September 6, 2019
Published: October 21, 2019

## REFERENCES


Akbari, H., Khalighinejad, B., Herrero, J.L., Mehta, A.D., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. Sci. Rep. *9*, 874.

Atiani, S., Elhilali, M., David, S.V., Fritz, J.B., and Shamma, S.A. (2009). Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. Neuron *61*, 467–480.

Atiani, S., David, S.V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S.A., and Fritz, J.B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. Neuron *82*, 486–499.

Bidelman, G.M., Moreno, S., and Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. Neuroimage *79*, 201–212.

Bizley, J.K., and Cohen, Y.E. (2013). The what, where and how of auditory-object perception. Nat. Rev. Neurosci. *14*, 693–707.

Bizley, J.K., Walker, K.M.M., Silverman, B.W., King, A.J., and Schnupp, J.W.H. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J. Neurosci. *29*, 2064–2075.

Bizley, J.K., Walker, K.M.M., Nodal, F.R., King, A.J., and Schnupp, J.W.H. (2013). Auditory cortex represents both pitch judgments and the corresponding acoustic cues. Curr. Biol. *23*, 620–625.

Bouchard, K.E.K.E., Mesgarani, N., Johnson, K., and Chang, E.F.E.F. (2013). Functional organization of human sensorimotor cortex for speech articulation. Nature *495*, 327–332.

Brodbeck, C., Hong, L.E., and Simon, J.Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. Curr. Biol. *28*, 3976–3983.

Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., and Lalor, E.C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Curr. Biol. *28*, 803–809.e3.

Brungart, D.S., Simpson, B.D., Ericson, M.A., and Scott, K.R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. J. Acoust. Soc. Am. *110*, 2527–2538.

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2010). Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. *13*, 1428–1432.

Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. *25*, 975–979.

Chi, T., Ru, P., and Shamma, S.A. (2005). Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. *118*, 887–906.

Clarke, S., and Morosan, P. (2012). Architecture, connectivity, and transmitter receptors of human auditory cortex. In The Human Auditory Cortex, D. Poeppel, T. Overath, A.N. Popper, and R.R. Fay, eds. (Springer), pp. 11–38.

Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences (Routledge).

David, S.V., Fritz, J.B., and Shamma, S.A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proc. Natl. Acad. Sci. USA *109*, 2144–2149.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. J. Neurosci. *37*, 6539–6557.

De Martino, F., Moerel, M., Xu, J., van de Moortele, P.F., Ugurbil, K., Goebel, R., Yacoub, E., and Formisano, E. (2015). High-resolution mapping of myeloarchitecture in vivo: Localization of auditory areas in the human brain. Cereb. Cortex *25*, 3394–3405.

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage *53*, 1–15.

Di Liberto, G.M., Wong, D., Melnik, G.A., and de Cheveigné, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage *196*, 237–247.

Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA *109*, 11854–11859.

Dykstra, A.R., Chan, A.M., Quinn, B.T., Zepeda, R., Keller, C.J., Cormier, J., Madsen, J.R., Eskandar, E.N., and Cash, S.S. (2012). Individualized localization and cortical surface-based registration of intracranial electrodes. Neuroimage *59*, 3563–3570.

Elhilali, M., Xiang, J., Shamma, S.A., and Simon, J.Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol. *7*, e1000129.

Fedorenko, E., Scott, T.L., Brunner, P., Coon, W.G., Pritchett, B., Schalk, G., and Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. Proc. Natl. Acad. Sci. USA *113*, E6256–E6262.

Fischl, B., Sereno, M.I., and Dale, A.M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage *9*, 195–207.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., et al. (2004). Automatically parcellating the human cerebral cortex. Cereb. Cortex *14*, 11–22.

Flinker, A., Doyle, W.K., Mehta, A.D., Devinsky, O., and Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. Nat. Hum. Behav. *3*, 393–405.

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., and Giraud, A.L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nat. Commun. *5*, 4694.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat. Neurosci. *6*, 1216–1223.

Gibson, J.J. (2014). The Ecological Approach to Visual Perception: Classic Edition (Psychology Press).

Groppe, D.M., Bickel, S., Dykstra, A.R., Wang, X., Mégevand, P., Mercier, M.R., Lado, F.A., Mehta, A.D., and Honey, C.J. (2017). iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data, *281*, pp. 40–48.

Hackett, T.A. (2008). Anatomical organization of the auditory cortex. J. Am. Acad. Audiol. *19*, 774–779.

Hackett, T.A., Preuss, T.M., and Kaas, J.H. (2001). Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. J. Comp. Neurol. *441*, 197–222.

Hamilton, L.S., Edwards, E., and Chang, E.F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr. Biol. *28*, 1860–1871.e4.

Han, C., O'Sullivan, J., Luo, Y., Herrero, J., Mehta, A.D., and Mesgarani, N. (2019). Speaker-independent auditory attention decoding without access to clean speech sources. Sci. Adv. *5*, eaav6134.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.

Hill, K.T., and Miller, L.M. (2010). Auditory attentional control and selection during cocktail party listening. Cereb. Cortex *20*, 583–590.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., and Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature *532*, 453–458.

Keller, C.J., Honey, C.J., Entz, L., Bickel, S., Groppe, D.M., Toth, E., Ulbert, I., Lado, F.A., and Mehta, A.D. (2014). Corticocortical evoked potentials reveal projectors and integrators in human brain networks. J. Neurosci. *34*, 9152–9163.

Kerlin, J.R., Shahin, A.J., and Miller, L.M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. *30*, 620–628.

Khalighinejad, B., Nagamine, T., Mehta, A., and Mesgarani, N. (2017). NAPLib: An open source toolbox for real-time and offline Neural Acoustic Processing.

In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference On (IEEE), pp. 846–850.

Khalighinejad, B., Herrero, J.L., Mehta, A.D., and Mesgarani, N. (2019). Adaptation of the human auditory cortex to changing background noise. Nat. Commun. *10*, 2509.

Khodagholy, D., Gelinas, J.N., Thesen, T., Doyle, W., Devinsky, O., Malliaras, G.G., and Buzsáki, G. (2015). NeuroGrid: recording action potentials from the surface of the brain. Nat. Neurosci. *18*, 310–315.

Kilian-Hütten, N., Valente, G., Vroomen, J., and Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. J. Neurosci. *31*, 1715–1720.

King, A.J., and Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat. Neurosci. *12*, 698–701.

Krishnan, L., Elhilali, M., and Shamma, S. (2014). Segregating complex sound sources through temporal coherence. PLoS Comput. Biol. *10*, e1003985.

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.

Lee, A.K.C., Larson, E., Maddox, R.K., and Shinn-Cunningham, B.G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. Hear. Res. *307*, 111–120.

Leonard, M.K., Bouchard, K.E., Tang, C., and Chang, E.F. (2015). Dynamic encoding of speech sequence probability in human temporal cortex. J. Neurosci. *35*, 7203–7214.

Leonard, M.K., Baud, M.O., Sjerps, M.J., and Chang, E.F. (2016). Perceptual restoration of masked speech in human cortex. Nat. Commun. *7*, 13619.

Linden, J.F., Liu, R.C., Sahani, M., Schreiner, C.E., and Merzenich, M.M. (2003). Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J. Neurophysiol. *90*, 2660–2675.

Luo, Y., and Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. *27*, 1256–1266.

Luo, Y., Chen, Z., and Mesgarani, N. (2018). Speaker-Independent Speech Separation With Deep Attractor Network. IEEE/ACM Trans. Audio Speech Lang. Process. *26*, 787–796.

Mesgarani, N., and Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature *485*, 233–236.

Mesgarani, N., David, S.V.S.V., Fritz, J.B.J.B., and Shamma, S.A.S.A. (2008). Phoneme representation and classification in primary auditory cortex. J. Acoust. Soc. Am. *123*, 899–909.

Mesgarani, N., Sivaram, G.S.V.S., Nemala, S.K., Elhilali, M., and Hermansky, H. (2009a). Discriminant spectrotemporal features for phoneme recognition. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2983–2986.

Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2009b). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J. Neurophysiol. *102*, 3329–3339.

Mesgarani, N., Fritz, J., and Shamma, S. (2010). A computational model of rapid task-related plasticity of auditory cortical receptive fields. J. Comput. Neurosci. *28*, 19–27.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science *343*, 1006–1010.

Miller, L.M., Escabí, M.A., Read, H.L., and Schreiner, C.E. (2001). Functional convergence of response properties in the auditory thalamocortical system. Neuron *32*, 151–160.

Miller, L.M., Escabí, M.A., Read, H.L., and Schreiner, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J. Neurophysiol. *87*, 516–527.

Moerel, M., De Martino, F., and Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. Front. Neurosci. *8*, 225.

Molenberghs, P., Mesulam, M.M., Peeters, R., and Vandenberghe, R.R.C. (2007). Remapping attentional priorities: differential contribution of superior parietal lobule and intraparietal sulcus. Cereb. Cortex *17*, 2703–2712.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., and Zilles, K. (2001). Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage *13*, 684–701.

Nelson, M.J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S.S., Naccache, L., Hale, J.T., Pallier, C., and Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. Proc. Natl. Acad. Sci. USA *114*, E3669–E3678.

Nourski, K.V. (2017). Auditory processing in the human cortex: An intracranial electrophysiology perspective. Laryngoscope Investig. Otolaryngol. *2*, 147–156.

Nourski, K.V., Steinschneider, M., Oya, H., Kawasaki, H., and Howard, M.A., 3rd (2015). Modulation of response patterns in human auditory cortex during a target detection task: an intracranial electrophysiology study. Int. J. Psychophysiol. *95*, 191–201.

Nourski, K.V., Steinschneider, M., Rhone, A.E., and Howard Iii, M.A. (2017). Intracranial Electrophysiology of Auditory Selective Attention Associated with Speech Classification Tasks. Front. Hum. Neurosci. *10*, 691.

Nourski, K.V., Steinschneider, M., Rhone, A.E., Kovach, C.K., Kawasaki, H., and Howard, M.A., 3rd (2019). Differential responses to spectrally degraded speech within human auditory cortex: An intracranial electrophysiology study. Hear. Res. *371*, 53–65.

O'Sullivan, J.A., Shamma, S.A., and Lalor, E.C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. J. Neurosci. *35*, 7256–7263.

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G.M.G.M., Sheth, S.A.S.A., Mehta, A.D.A.D., Mesgarani, N., et al. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. J. Neural Eng. *14*, 056001.

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J.P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. Cereb. Cortex *17*, 2251–2257.

Papademetris, X., Jackowski, M.P., Rajeevan, N., DiStasio, M., Okuda, H., Constable, R.T., and Staib, L.H. (2006). BioImage Suite: An integrated medical image analysis suite: An update. Insight J. *2006*, 209.

Patel, P., Long, L.K., Herrero, J.L., Mehta, A.D., and Mesgarani, N. (2018). Joint Representation of Spatial and Phonetic Features in the Human Core Auditory Cortex. Cell Rep. *24*, 2051–2062.e2.

Petkov, C.I., Kang, X., Alho, K., Bertrand, O., Yund, E.W., and Woods, D.L. (2004). Attentional modulation of human auditory cortex. Nat. Neurosci. *7*, 658–663.

Power, A.J., Foxe, J.J., Forde, E.J., Reilly, R.B., and Lalor, E.C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. Eur. J. Neurosci. *35*, 1497–1503.

Puschmann, S., Baillet, S., and Zatorre, R.J. (2018). Musicians at the cocktail party: neural substrates of musical training during selective listening in multi-speaker situations. Cereb. Cortex.

Puvvada, K.C., and Simon, J.Z. (2017). Cortical Representations of Speech in a Multi-talker Auditory Scene. J. Neurosci. *37*, 0938–17.

Rademacher, J., Caviness, V.S., Jr., Steinmetz, H., and Galaburda, A.M. (1993). Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology. Cereb. Cortex *3*, 313–329.

Rasmussen, G.L. (1964). Anatomic relationships of the ascending and descending auditory systems. Neurol. Asp. Audit. Vestib. Disord. *1*, 5–19.

Rauschecker, J.P. (1997). Processing of complex sounds in the auditory cortex of cat, monkey, and man. Acta Otolaryngol. Suppl. *532*, 34–38.

Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. *12*, 718–724.

Salmi, J., Rinne, T., Koistinen, S., Salonen, O., and Alho, K. (2009). Brain networks of bottom-up triggered and top-down controlled shifting of auditory attention. Brain Res. *1286*, 155–164.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. *10*, e1003412.

Shamma, S. (2008). On the emergence and awareness of auditory objects. PLoS Biol. *6*, e155.

Shamma, S.A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. Trends Neurosci. *34*, 114–123.

Shinn-Cunningham, B.G. (2008). Object-based auditory and visual attention. Trends Cogn. Sci. *12*, 182–186.

Slee, S.J., and David, S.V. (2015). Rapid task-related plasticity of spectrotemporal receptive fields in the auditory midbrain. J. Neurosci. *35*, 13090–13102.

Steinschneider, M., Nourski, K.V., and Fishman, Y.I. (2013). Representation of speech in human auditory cortex: is it special? Hear. Res. *305*, 57–73.

Steinschneider, M., Nourski, K.V., Rhone, A.E., Kawasaki, H., Oya, H., and Howard, M.A., 3rd (2014). Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. Front. Neurosci. *8*, 240.

Teki, S., Barascud, N., Picard, S., Payne, C., Griffiths, T.D., and Chait, M. (2016). Neural correlates of auditory figure-ground segregation based on temporal coherence. Cereb. Cortex *26*, 3669–3680.

Thakur, C.S., Wang, R.M., Afshar, S., Hamilton, T.J., Tapson, J.C., Shamma, S.A., and van Schaik, A. (2015). Sound stream segregation: a neuromorphic approach to solve the "cocktail party problem" in real-time. Front. Neurosci. *9*, 309.

Theunissen, F.E., Sen, K., and Doupe, A.J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J. Neurosci. *20*, 2315–2331.

Walker, K.M.M., Bizley, J.K., King, A.J., and Schnupp, J.W.H. (2011). Multiplexed and robust representations of sound features in auditory cortex. J. Neurosci. *31*, 14565–14576.

Webster, D.B., and Fay, R.R. (2013). The Mammalian Auditory Pathway: Neuroanatomy (Springer Science & Business Media).

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron *77*, 980–991.

# STAR★METHODS

## LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Nima Mesgarani (nima@ee.columbia.edu). This study did not generate new unique reagents.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Human Subjects

Eight subjects who were undergoing clinical treatment for epilepsy participated in this study. All subjects gave their written informed consent to participate in research. Five subjects were located at North Shore University Hospital (NSUH), and 3 subjects were located at Columbia University Medical Center (CUMC). All research protocols were approved and monitored by the institutional review board at the Feinstein Institute for Medical Research and Columbia University Medical Center (CUMC) and informed written consent to participate in research studies was obtained from each subject before implantation of electrodes. Two subjects were implanted with high density subdural electrode arrays over the left (language dominant) temporal lobe with coverage of STG, and one of those subjects also had a depth electrode implanted in the left auditory cortex with coverage of HG. The remaining 6 subjects had depth electrodes implanted bilaterally, with varying amounts of coverage over the left and right auditory cortices for each subject (Figure 1A).

### Stimuli and Experiments

Each subject participated in the following experiments for this study: a single-talker and multi-talker experiment. The single-talker experiment was used as a control. Each subject listened to 4 stories read by a male speaker and female speaker (hereafter referred to as Spk1 and Spk2, respectively) for a total of 8 stories (4 stories twice). Each story lasted approximately 3.5 minutes. Both Spk1 and Spk2 were native American English speakers and were recorded in house. The average F0 of Spk1 and Spk2 was 65Hz and 175Hz, respectively. The stories were intermittently paused, and the subject was instructed to repeat the last sentence to ensure the attentional engagement of each subject. For the multi-talker experiment, subjects were presented with a mixture of the same female and male speakers (Spk1 and Spk2), with no spatial separation between them. The acoustic waveform of each speaker was matched to have the same root mean squared (RMS) intensity. All stimuli were presented using a single Bose® SoundLink® Mini 2 speaker situated directly in front of the subject. The sound level was adjusted for each subject to be at a comfortable level.

The multi-talker experiment was divided into 4 blocks. Before each block, the subject was instructed to focus their attention on one speaker and to ignore the other. All subjects began the experiment by attending to the male speaker and switched their attention to the alternate speaker on each subsequent block. The story was intermittently paused, and the subjects were asked to repeat the last sentence of the attended speaker to ensure that the subjects were engaged in the task. The stories were paused on average every 20 s (min 9 s, max 30 s). The locations of the pauses were predetermined and were the same for all subjects, but the subjects were unaware of when the pauses would occur. In total, there were 11 minutes and 37 s of audio presented to each subject during the multi-talker experiment. The single-talker experiment lasted twice as long as each subject was required to listen to each story read by each speaker independently.

## METHOD DETAILS

### Data Preprocessing and Hardware

The subjects at NSUH were recorded using Tucker Davis Technologies (TDT®) hardware and sampled at 2441 Hz. One subject at CUMC was recorded using Xltek® hardware and sampled at 500 Hz, and the other 2 subjects at CUMC were recorded using Blackrock® hardware and sampled at 3 kHz. All further processing steps were performed offline. All filters were designed using MATLAB's® Filter Design Toolbox and were used in both forward and backward directions to remove phase distortion. The TDT and Blackrock data were resampled to 500 Hz. A $1^{st}$-order Butterworth high-pass filter with a cut-off frequency at 1 Hz was used to remove DC drift. Data were subsequently re-referenced using a local scheme whereby each electrode was referenced relative to its nearest neighbors. Line noise at 60 Hz and its harmonics (up to 240 Hz) were removed using $2^{nd}$ order IIR notch filters with a bandwidth of 1 Hz. A period of silence lasting 1 minute was recorded before the single-talker and multi-talker experiments, and the corresponding data were normalized by subtracting the mean and dividing by the standard deviation of this pre-stimulus period.

Then, data were filtered into the high-gamma band (70-150 Hz); the envelope of this band is modulated by speech. To obtain the envelope of this broad band, we first filtered the data into 8 frequency bands between 70 and 150 Hz, each with a bandwidth of 10 Hz, using Chebyshev Type 2 filters. Then, the envelope of each band was obtained by taking the absolute value of the Hilbert transform. We took the average of all 8 frequency bands as the final envelope. This method is commonly used in neuroscience research(Bouchard et al., 2013). Electrodes were tested for speech responsiveness by calculating the effect size (Cohen's D) between the

distributions of the responses during speech and silence (instantaneous envelope of the high gamma band at each time point). Electrodes with an effect size greater than 0.2 (considered a small but significant effect size) were retained for further analysis.

### Acoustic Spectrum of Speakers

The spectrograms were first z-scored and then filtered using the NSL toolbox(Chi et al., 2005), specifically using the static cortical representation (aud2cors) to obtain the average acoustic spectrum of each speaker. This process provided a more defined representation of the harmonic structure of both speakers. The spectrograms were sampled at 100Hz, and split into 50 frequency bands logarithmically spaced between 50Hz and 8kHz.

### STRFs and Stimulus Reconstruction

STRFs and stimulus reconstruction decoders were calculated using custom code to implement ridge regression. K-fold cross-validation was used to select a ridge parameter that would optimally predict neural data in the case of a STRF or optimally reconstruct spectrograms in the case of stimulus reconstruction. To estimate the response latency of an electrode, we took the peak magnitude of its STRF after averaging across frequency. For reconstruction, we used only electrodes in HG whose AMI was less than the threshold previously established for statistical significance (i.e., AMI < 0.15). Only 4 electrodes were rejected using this criterion.

### Predicting Speaker Selectivity and Attentional Modulation

Before predicting the SSI of each site, we removed the temporal dimension of their STRFs by obtaining their 1$^{st}$ PC in the spectral dimension. Therefore, we will abbreviate these STRFs to SRFs. Next, we used ridge regression to find a set of weights that would predict the SSI for each site using its SRF (see Figure S1). Ten-fold cross-validation was used to optimize the ridge parameter.

### Mapping HG to STG

Mappings between HG and STG were calculated in the same manner as the STRFs (i.e., k-fold ridge regression). However, only causal time lags were used (0-400 ms). In addition, only STG electrodes from the two subjects with high density grids were used. This requirement was used to prevent predictions with spuriously large correlations with the actual data due to shared noise between contacts on the same electrode arrays. That is, we used data from the depth electrodes to predict the data on the grid electrodes.

### Temporal Coherence

Because every retained HG electrode responded significantly to the presence of speech, this introduced spuriously large correlations across all electrodes. To remove this confound and focus more on what was discriminative between electrodes, we subtracted the first PC from the neural responses. Then, we found the correlation between all HG electrodes to obtain a 2D pairwise correlation matrix. To relate this 2D matrix to the 1D array of speaker-selectivity indices, we performed PCA on this matrix. The correlations between speaker selectivity and the first 3 PCs are 0.76, 0.3 and 0.35. The correlation in the results section (Figure 8B) is between speaker selectivity and the sum of the first 3 PCs. The matrix shown (Figure 8A) is the projection of the first 3 PCs onto the 2D pairwise correlation matrix.

### Transformation of Electrode Locations onto an Average Brain

The electrodes were first mapped onto the brain of each subject using co-registration by iELVis (Groppe et al., 2017) followed by their identification on the post-implantation CT scan using BioImage Suite(Papademetris et al., 2006). To obtain the anatomical location labels of these electrodes, we used Freesurfer's automated cortical parcellation(Dykstra et al., 2012; Fischl et al., 1999, 2004) by the Destrieux brain atlas(Destrieux et al., 2010). These labels were closely inspected by the neurosurgeons using the subject's co-registered post-implant MRI. We plotted the electrodes on the average Freesurfer brain template.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Speaker-Selectivity Index (SSI)

The SSI was calculated as the effect size (Cohen's D) of the difference in the magnitude of the responses to each speaker in the single-talker condition. Figure 2A shows histograms of the responses to Spk1 and Spk2 in the single-talker condition for 2 example electrodes. The responses were normalized by concatenating the responses to each speaker together, and then z-scoring them. This ensured that any difference in response magnitude to either speaker would be maintained in the normalized representation.

### Attentional Modulation Index (AMI)

A chance level for the AMI was obtained by randomly shuffling the temporal order of the neural data and calculating the AMI per electrode as follows:

$$AMI = corr(Spk1_{attend}, Spk1_{alone}) - corr(Spk1_{attend}, Spk2_{alone}) + corr(Spk2_{attend}, Spk2_{alone}) - corr(Spk2_{attend}, Spk1_{alone})$$

Where *SpkX* refers to the response to speaker *X* either in the single-talker condition (alone) or when they are attended in the multi-talker condition (attend).

This calculation resulted in a null distribution of the AMI. As expected, the chance level of the AMI was zero (mean of the null distribution). To determine what could be considered an AMI significantly above chance, we used three times the standard deviation of the null distribution, which corresponds to an AMI of 0.15. Figure S3 shows a linear correlation between speech responsiveness (effect size: speech versus silence) and AMI in STG ($r = 0.71$, $p < 0.001$) but not in HG ($r = 0$). This result was observed probably because our measure of attention is based on the correlation between the multi- and single-talker responses and is affected by the signal-to-noise ratio (SNR) of the recording at each electrode.

## DATA AND CODE AVAILABILITY

There are restrictions to the availability of dataset due to the protection of human subjects who participated in this study. The data that support the findings of this study are available upon request from the corresponding author [NM]. The codes for pre-processing the ECoG signals and calculating the high-gamma envelope are available at http://naplab.ee.columbia.edu/naplib.html (Khalighinejad et al., 2017).