Article

Current Biology

Interaction of bottom-up and top-down neural mechanisms in spatial multi-talker speech perception

Highlights

- We studied the encoding of spatially separated multi-talkers in the human auditory cortex
- Contralateral tuning to speech features sharpens by the spatial separation of talkers
- Attended talker's location and voice modulate different components of neural response
- Attention refines a separable bottom-up representation of the talker's voice and location

Authors

Prachi Patel, Kiki van der Heijden, Stephan Bickel, Jose L. Herrero, Ashesh D. Mehta, Nima Mesgarani

Correspondence

nima@ee.columbia.edu

In brief

Direct neural recording from the human auditory cortex revealed an interaction between bottom-up and top-down neural processes that underlie spatial multi-talker speech perception. The voice and location of the attended talker differentially modulated neural response, thus further refining a separable pre-attentive neural representation of the talker.

Patel et al., 2022, Current Biology *32*, 3971–3986
September 26, 2022
https://doi.org/10.1016/j.cub.2022.07.047

CellPress

Current Biology

CellPress

Article

Interaction of bottom-up and top-down neural mechanisms in spatial multi-talker speech perception

Prachi Patel,[1,2] Kiki van der Heijden,[1,3,6] Stephan Bickel,[4,5] Jose L. Herrero,[4,5] Ashesh D. Mehta,[4,5] and Nima Mesgarani[1,2,7,8,*]

[1]Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA
[2]Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
[3]Donders Institute for Brain Cognition and Behavior, Radboud University, Nijmegen, the Netherlands
[4]Hofstra Northwell School of Medicine, New York, NY 11549, USA
[5]The Feinstein Institute for Medical Research, New York, NY 11030, USA
[6]Maastricht Centre for Systems Biology, Faculty of Science and Engineering, Maastricht University, Maastricht, the Netherlands
[7]Twitter: @NimaMesgarani
[8]Lead contact
*Correspondence: nima@ee.columbia.edu
https://doi.org/10.1016/j.cub.2022.07.047

SUMMARY

How the human auditory cortex represents spatially separated simultaneous talkers and how talkers' locations and voices modulate the neural representations of attended and unattended speech are unclear. Here, we measured the neural responses from electrodes implanted in neurosurgical patients as they performed single-talker and multi-talker speech perception tasks. We found that spatial separation between talkers caused a preferential encoding of the contralateral speech in Heschl's gyrus (HG), planum temporale (PT), and superior temporal gyrus (STG). Location and spectrotemporal features were encoded in different aspects of the neural response. Specifically, the talker's location changed the mean response level, whereas the talker's spectrotemporal features altered the variation of response around response's baseline. These components were differentially modulated by the attended talker's voice or location, which improved the population decoding of attended speech features. Attentional modulation due to the talker's voice only appeared in the auditory areas with longer latencies, but attentional modulation due to location was present throughout. Our results show that spatial multi-talker speech perception relies upon a separable pre-attentive neural representation, which could be further tuned by top-down attention to the location and voice of the talker.

INTRODUCTION

Humans can attend to a single talker in acoustically complex multi-talker environments, particularly when the talkers are separated in space. Successful perception of a target talker's speech requires one to identify this speech and separate it from the background, using both spectrotemporal and spatial information of the talker. The cognitive mechanisms underlying this challenging task are thought to rely on complex interactions between pre-attentive (bottom-up) and attentive (top-down) processing of the acoustic scene. Top-down processing is said to happen at least at two levels: the higher-order level of attention control involving the prefrontal cortex[1,2] and the relatively lower-order attention selection at the level of the auditory cortex.[3,4] This study focuses on the latter. Past studies exploring the mechanism of top-down attention selection have suggested that simultaneous sounds activate well-separated neural populations in the cortex and that this process is largely bottom-up,[5,6] whereas attention (top-down selection) merely selects one of these streams for further processing.[7,8] Other studies have argued that stream segregation is largely driven and shaped by attentional focus,[9,10] either by enhancing target-specific feature representation in a bottom-up manner[8] or by modulating the temporal coherence of neural populations, thus binding various features of the attended stream.[9,11] In our previous study, we explored the neural encoding of a single talker in space by characterizing the joint representation of location and spectrotemporal and phonetic features in the human auditory cortex.[12] In the present study, we examined how the neural representation of location, spectrotemporal features, and phonetic features vary when the subject attends to one of the talkers in a specific location in multi-talker speech perception. Specifically, we studied how the presence of an additional talker in space changes the neural encoding of each talker (bottom-up), as well as how the attentional focus (top-down) modifies the neural representation of a talker's location and spectrotemporal and phonetic features. In the subsequent paragraphs, we discuss our current understanding and remaining questions regarding the bottom-up and top-down neural representation of spatial auditory objects.

Our knowledge of how spatial separation between multiple streams modifies the neural encoding of the individual streams (in a bottom-up manner) comes partly from non-human animals. Many of these studies have attempted to decipher the spatial mechanisms of auditory scene analysis during passive listening[13,14] or under anesthesia.[15–17] In delineating the bottom-up pre-attentive neural representations, both spatial and spectrotemporal cues are important, as are their interactions with each other.[18] For spatial cues, responses from the auditory cortex of anesthetized animals have demonstrated that single neurons exhibit bottom-up stream segregation by narrowing their tuning to the location of one of the multiple simultaneous spatial sound sources.[19,20] However, it is unknown whether the processes of passive spatial stream segregation and active sound localization occur within common cortical areas,[19,20] or whether there are distinct cortical areas dedicated to bottom-up segregation and top-down sound localization.[21–24] This is important as previous studies have shown that the behavioral state of listeners affects the spatial selectivity of neural responses.[25,26] For spectrotemporal cues, engagement in a nonspatial multistream task as opposed to passive listening has been shown to modulate spectrotemporal receptive fields (STRFs) in the mammalian auditory cortex, thereby enhancing the response to a particular auditory object of attention,[27,28] with stronger enhancement in nonprimary than in primary areas.[29] What remains unclear is the underlying mechanisms of how neural populations in the human brain are distinguished by spatial sensitivity and thus how they integrate spatial and spectral cues for segregation to identify sources that could further accomplish target selection.[30]

In humans, the effects of top-down attention in changing the representation of talkers' features like their location, voice, and phonemes have been studied extensively. For example, attention to a talker's voice in spatially overlapping multi-talker speech has been shown to modulate spectrotemporal tuning of neural responses to enhance encoding of an attended talker.[3,4,31,32] Moreover, studies have confirmed a top-down effect of attention to location, which decreases alpha band (8–12 Hz) power in the brain hemisphere contralateral to the location to which the subject attends.[33,34] However, the neural representation of naturalistic multi-talker speech with spatial separation between talkers remains largely unexplored. Unlike the studies of co-located multi-talker speech,[3,4,31,32] spatial separation between talkers raises new questions that have been hard to study due to difficulties in measuring the neural responses within deeper cortical structures where the bottom-up effects are shown to be predominant in animal models.[19,20] As a result, it is unclear how the spatial separation between multiple talkers changes the representation of spectrotemporal, phonetic, and spatial features in the human auditory cortex (stimulus-driven or bottom-up). It is further unknown how the attended talker's location and voice are represented jointly in the human auditory cortex (top-down attention selection). Finally, it remains unclear how these bottom-up and top-down signals interact to enable a successful decoding of both the talker's message and location in space.

Here, we measured neural responses using intracranial electrodes implanted in the auditory cortex of neurosurgical patients while they performed single-talker and multi-talker speech perception tasks with and without spatial separation of talkers. We aimed to characterize and separate the neural encoding of talkers' speech and location and to study the interaction of bottom-up (stimulus-driven) effects and top-down (attention selection) neural effects on the encoding of these features in multi-talker listening conditions. Our findings reveal that spatial separation between simultaneous talkers causes a preferential encoding of the speech features of the contralateral talker over the ipsilateral talker. Moreover, we find that the talker's speech and location are encoded in different aspects of the neural responses and that top-down selective attention specifically modulates these aspects. Together, these findings advance our understanding of how the human auditory cortex represents and processes spatially separated simultaneous talkers to enable the perception of spatial multi-talker speech.

RESULTS

We recorded from high-density depth electrodes-stereotactic electroencephalography (sEEG) implanted in the auditory cortex of seven patients with epilepsy. Three subjects had bilateral implants and four subjects had implants only in the right hemisphere. The electrodes provided partial coverage of several areas, including in Heschl's gyrus (HG) and sulcus, the planum temporale (PT), and the superior temporal gyrus (STG) and sulcus (Figure 1A).[35] The subjects listened to speech stimuli arriving from two locations in the azimuthal plane (−45 and 45 degrees) simulated using the head-related transfer function (HRTF) of an average-sized head (STAR Methods) and delivered via headphones. The subjects performed three speech perception tasks: they performed a single-talker task (Figure 1B) where they listened to speech uttered by a male or female talker from the left (−45 degrees) or the right (45 degrees). The location and the talker (male or female) changed randomly after each trial, which was 5 s long on average. The subjects also performed a spatially separated multi-talker task (Figure 2A) where the subjects listened to simultaneous speech of a male and a female talker separated in space. Finally, the subjects performed a spatially overlapping multi-talker task (Figure 3A) where the subjects listened to the simultaneous male and female talker collocated on either side of the space. The speech in each of these three experimental conditions was 8 min long. For the single-talker task, the subjects were asked to report the location and repeat the content of the last sentence uttered by the talker when speech was paused at random intervals. For multi-talker tasks, the subjects were instructed to attend to either the male or the female talker throughout a given experiment block while the location of the attended talker randomly switched between the two locations after each sentence (totaling in two blocks: attend-male and attend-female). The order of the tasks performed was as follows: spatially separated multi-talker, spatially overlapping multi-talker, and then single talker. All subjects were fluent speakers of American English. We asked the subjects to report the location and the content of the last sentence spoken by the attended talker at random intervals. Subjects performed with a mean accuracy of 90.2% for a single talker, 93.8% for multiple spatially separated talkers, and 89.3% for multiple collocated talkers, thus performing the task equally well in single- and multi-talker conditions (Figure S1A).

We used the envelope of the high gamma band (70–150 Hz) as our neural response measure; this measure has been shown to

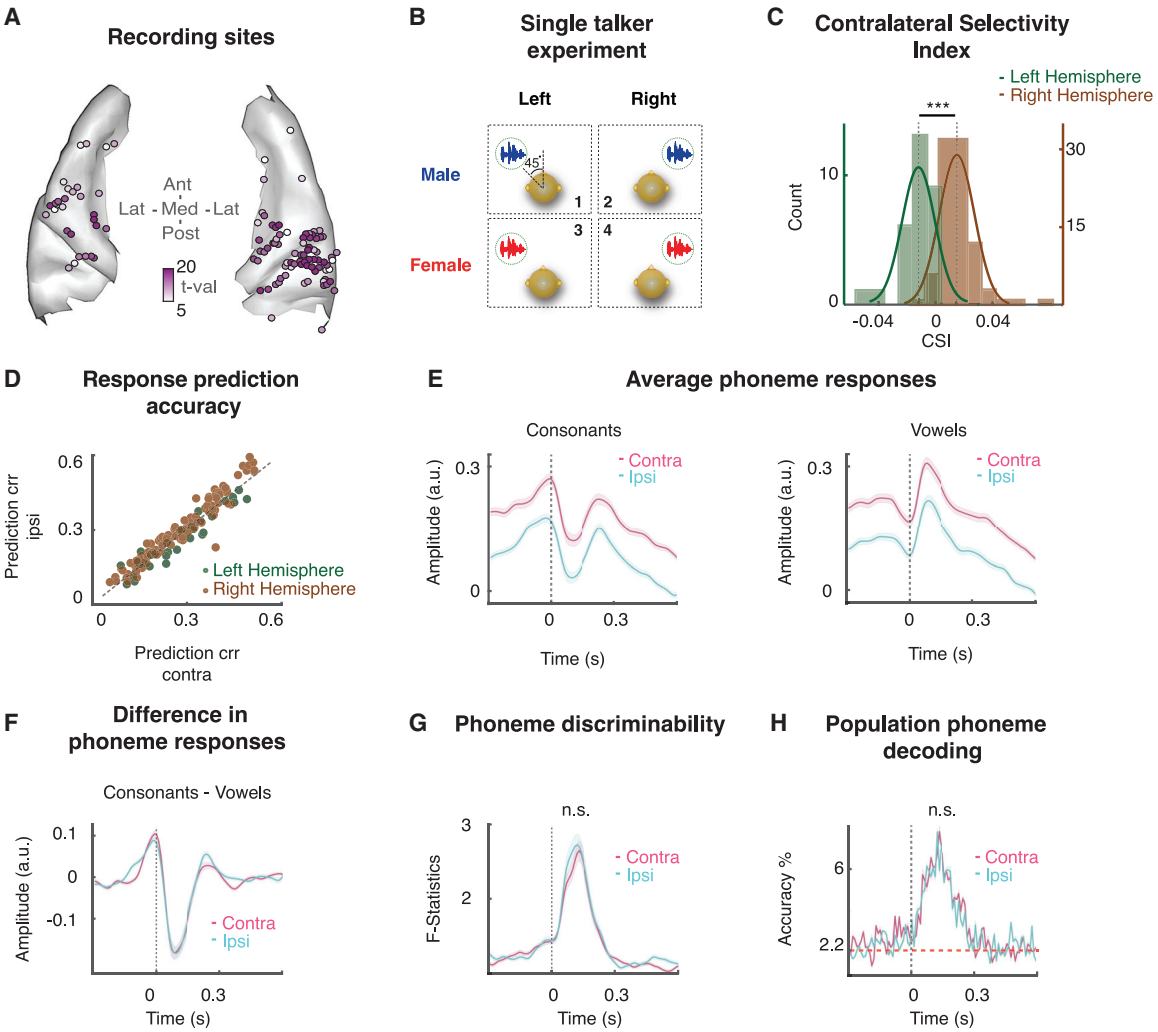

Figure 1. Neural encoding of location and speech in the single-talker condition

(A) Speech-responsive electrodes from all subjects shown on an ICBM152 average brain. Color saturation indicates the speech versus silence t value for each electrode.

(B) Task schematic. Subjects are presented with speech uttered from two angles in the horizontal plane (45 and −45 degrees) and two talkers (male and female).

(C) Histograms of contralateral selectivity index (CSI) from mean high gamma responses of individual electrodes for the left and the right brain hemispheres ($p < 0.001$, unpaired t test).

(D) Scatterplot of STRF prediction correlations for single electrodes for speech from contralateral (x axis) and ipsilateral (y axis) locations. Electrodes are colored by brain hemisphere. See also Figure S1.

(E) Average high gamma responses to consonants and vowels for speech from contralateral (pink) and ipsilateral (blue) locations.

(F) Average high gamma responses to consonants minus vowels from the contralateral (pink) and ipsilateral (blue) locations.

(G) Average F-statistic from single electrodes for separation of 46 individual phonemes from the contralateral (pink) and ipsilateral (blue) locations ($p > 0.05$, paired t test).

(H) Population phoneme decoding accuracy (y axis) for 46 individual phonemes from the contralateral (pink) and ipsilateral (blue) locations ($p > 0.05$, paired t test).

directly reflect the average firing of nearby neurons.[36,37] We then identified the electrodes that had a significant response to speech and restricted our analysis to this subset of electrodes (STAR Methods). This resulted in a scope of 119 electrodes, including 32 electrodes in Heschl's gyrus and sulcus, 59 in the superior temporal gyrus and sulcus, and 28 in the planum temporale (Figure 1A).[35] We have organized the figures according to the main findings of our study: Figure 1 illustrates the baseline by examining the encoding of a single spatial talker prior to adding a second talker to the acoustic scene; Figures 2 and 3 explore the effect of bottom-up stimulus-driven processing; Figures 4 and 5 examine the effect of top-down attention on the representation of talker's location, spectrotemporal features, and phonetic features; Figures 6 and 7 show the anatomical organization of these neural effects.

Neural encoding of the talker's location and speech in a single-talker acoustic condition

To study the neural encoding of the talker's voice and location in the single-talker condition, we tested how these two talker

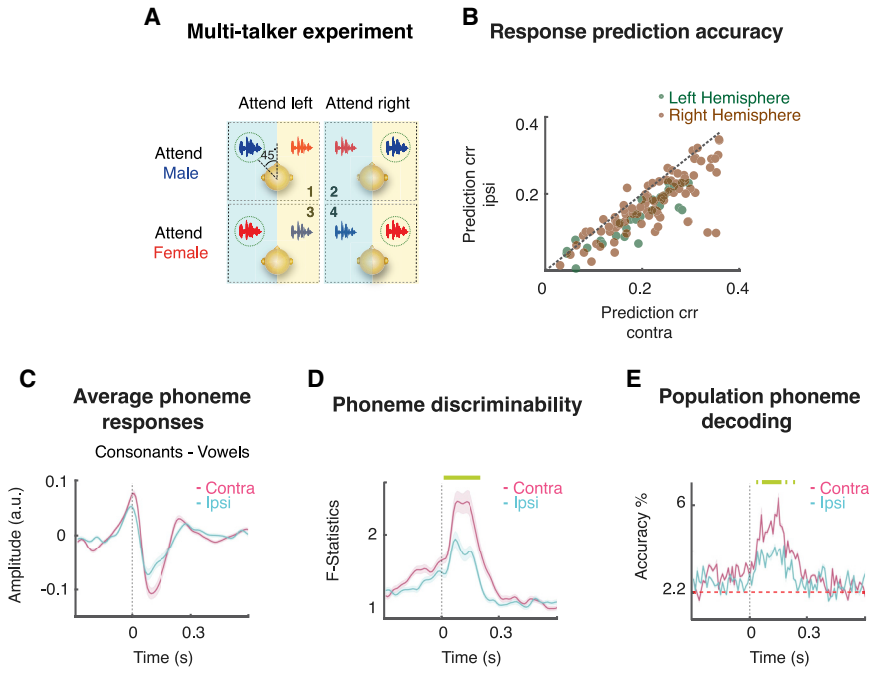

Figure 2. Bottom-up spatial multi-talker condition: Neural spatial tuning sharpens to the speech of the contralateral talker

(A) Task schematic. Simultaneous spatially separated talkers in four conditions (1–4) for combinations of attention to two locations and two talkers; comparison is made between neural encoding of all the speech (attended and unattended combined) from left (blue) versus all the speech (attended and unattended combined) from right (yellow).

(B) Scatterplot of STRF prediction correlations for single electrodes for speech from contralateral (x axis) and ipsilateral (y axis) locations. Electrodes are colored by brain hemisphere. See also Figure S1.

(C) Average high gamma responses to consonants minus vowels from the two locations, with contralateral in pink and ipsilateral in blue.

(D) Average F-statistic from single electrodes for the separation of 46 individual phonemes from the contralateral (pink) and ipsilateral (blue) locations ($p < 0.001$, paired t test).

(E) Population phoneme decoding accuracy (y axis) for 46 individual phonemes from the contralateral (pink) and ipsilateral (blue) locations ($p < 0.001$, paired t test).

features are represented in the neural responses. First, we examined the effect of the talker's location on the neural response. We measured the baseline neural activity to the speech from each talker location by measuring the mean response level (MRL). The MRL for each location was calculated by averaging the neural responses to all speech tokens from that location for each electrode. We then quantified location selectivity by defining the contralateral selectivity index (CSI) for each electrode as the difference between the mean neural response to the left location and the right location normalized by their sum[12] (STAR Methods). Figure 1C shows a histogram of CSI (x axis) of all 119 electrodes separated by brain hemispheres (brown: right hemisphere; green: left hemisphere). Electrodes in the right hemisphere show positive CSI, meaning that they have a higher MRL for the left (contralateral) than for the right (ipsilateral), and vice versa for electrodes in the left hemisphere. The absolute difference between means is 0.03 (Figure 1C, $p < 0.001$, unpaired t test). This finding is consistent with our past study, which showed a higher MRL in response to speech arriving from the contralateral location than the ipsilateral location.[12]

To examine whether this MRL shift due to the talker's location impacts the encoding of spectrotemporal features of the talker's voice, we examined and compared the prediction accuracy of the STRF model (STAR Methods) for the two stimulus locations. For the subsequent analysis, we normalized the electrode data to a zero mean and unit standard deviation. For unbiased comparison, we calculated STRFs from a single-channel spectrogram with no spatial information and compared STRF predictions for speech from the contralateral and ipsilateral locations. Figure 1D shows a scatterplot of STRF prediction coefficients for speech from contralateral (x axis) and ipsilateral (y axis) locations. Each dot in Figure 1D represents a single electrode and is colored according to its hemispheric location in the brain. The lack of difference between the prediction coefficients from the contralateral and ipsilateral locations ($p > 0.05$, permutation test) shows that in the single-talker condition, the encoding accuracy of spectrotemporal acoustic features does not depend on the location of the talker, despite the variation in the electrode's MRL.

While spectrotemporal features accurately predict the representation of sound in the auditory nerve,[38] the human auditory cortex is specialized for speech processing[39] and is shown to encode phonetic features of speech.[40,41] To study the effect of talker location on phoneme encoding, we transcribed and segmented the continuous speech stimulus into two groups, namely consonants and vowels,[42] and averaged the neural response over all the instances of each group. Figure 1E qualitatively shows the average response to consonants and vowels when the talker was on the opposite side (contralateral) or the same side (ipsilateral) relative to the hemispheric location of the electrodes. Consistent with Figure 1C, we observed that the MRL varied with the location of the talker and was higher when the talker was on the contralateral side relative to the electrodes (Figure 1E). Subtracting the average vowel response from the average consonant response separately for each of the two locations eliminates the mean response and highlights the similarity of the neural response (other than the mean) to speech irrespective of the location (Figure 1F). We quantified this effect for individual electrodes by measuring the phoneme discriminability using the F-statistic[43] to calculate the separation between neural responses to 46 individual phonemes. Figure 1G shows an overlapping average neural F-statistic over time (0 is the phoneme onset) for individual electrodes from the contralateral (pink) and ipsilateral (blue) locations ($p > 0.05$, paired t test). In addition to the electrode-level analysis in Figure 1G, we also quantified the population-level (combined electrodes) phoneme decoding accuracy as a function of location. We used a linear

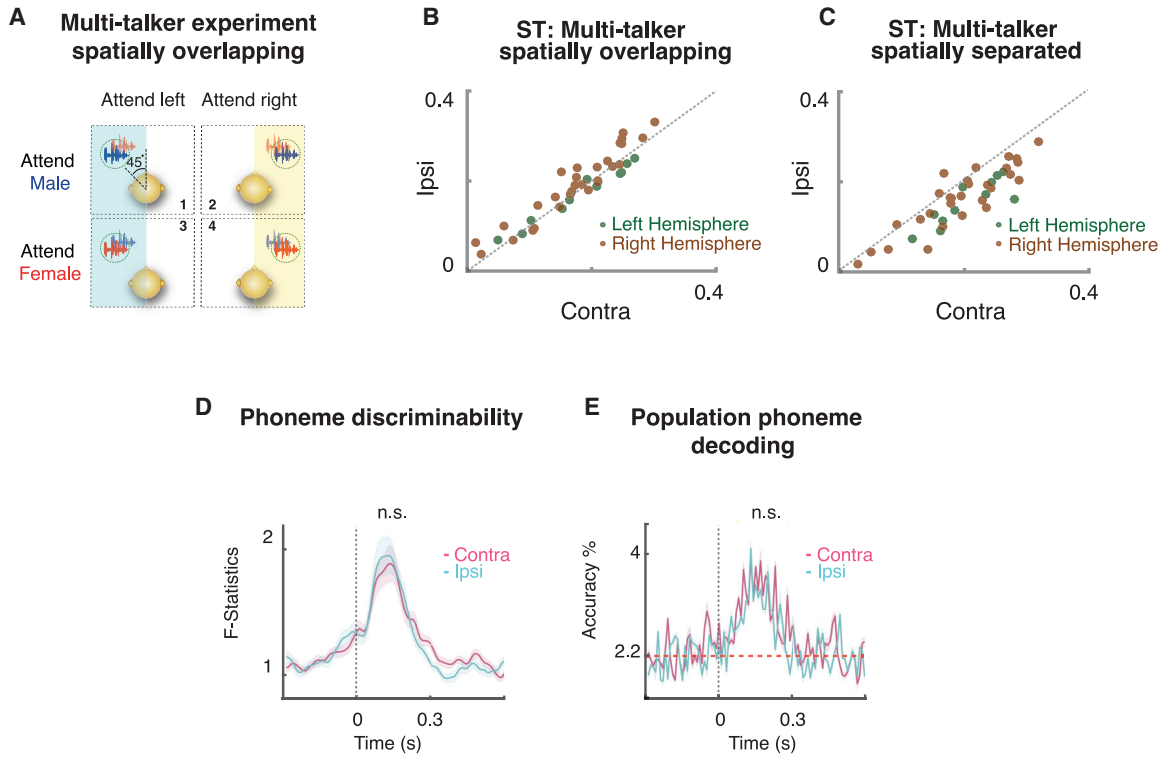

Figure 3. Neural spatial tuning sharpness disappears for spatially overlapping talkers

(A) Task schematic. Simultaneous spatially overlapping talkers in four conditions (1–4) for combinations of attention to two locations and two talkers; comparison between the encoding of speech from the left (blue) and right (yellow).

(B) Scatterplot of STRF prediction correlations for single electrodes for speech from contralateral (x axis) and ipsilateral (y axis) locations in spatially overlapping multi-talker case for 2 subjects. Electrodes are colored by brain hemisphere.

(C) Scatterplot of STRF prediction correlations for single electrodes for speech from contralateral (x axis) and ipsilateral (y axis) locations in spatially separated multi-talker case for 2 subjects. Electrodes are colored by brain hemisphere. See also Figure S1.

(D) Average F-statistic from single electrodes for the separation of 46 individual phonemes from the contralateral (pink) and ipsilateral (blue) locations ($p > 0.05$, paired t test).

(E) Population phoneme decoding accuracy (y axis) for 46 individual phonemes from the contralateral (pink) and ipsilateral (blue) locations ($p > 0.05$, paired t test).

regularized least squares (RLS) classifier[44] to decode 46 individual phonemes from the contralateral location and the ipsilateral location (Figure 1H; STAR Methods). Similar to the individual electrode analysis (Figure 1G), the population neural decoding also shows similar decoding accuracy from either location ($p > 0.05$, paired t test). To summarize, the lack of significant separation between the plots in Figures 1D, 1G, and 1H indicates that in the single-talker condition, the neural encoding of spectrotemporal and phoneme features is unaffected by the location of the talker in space, although the location of the talker changes the MRL of the response (Figures 1C and 1E). Next, we examine how the encoding of speech and location changes in multi-talker scenarios.

Enhanced bottom-up representation of contralateral speech in multi-talker conditions

We showed that the location of the talker in the single-talker condition does not change the spectrotemporal and phonetic encoding of speech, but it changes the MRL. On that precedent, we then tested whether this effect also holds in multi-talker acoustic conditions. In other words, we tested whether simultaneous speech from contralateral and ipsilateral locations is encoded with similar accuracy in the auditory cortex. This analysis aimed to examine the bottom-up encoding of speech features, but since the task always required the subject to focus on a specific talker, we needed to control for the possible modulatory effects of attention. We therefore averaged out the attentional effects by combining across all attended and unattended trials from the contralateral side and compared them with all combined attended and unattended trials from the ipsilateral side. For example, we compared the response of an electrode in the left brain hemisphere between ipsilateral locations (Figure 2A, in blue, attended and unattended combined on the ipsilateral side) and contralateral locations (Figure 2A, in yellow, attend and unattended combined on the contralateral side).

We first examined the encoding accuracy of spectrotemporal features of speech from contralateral and ipsilateral talkers using STRFs. For a given electrode, we measured nonspatial STRFs (calculated using a single-channel spectrogram without spatial features) for contralateral speech and ipsilateral speech separately and compared the prediction accuracy of the two. Figure 2B shows the correlation between actual and predicted responses from STRFs for contralateral and ipsilateral talkers. In contrast to the single-talker case (Figure 1D), we found that

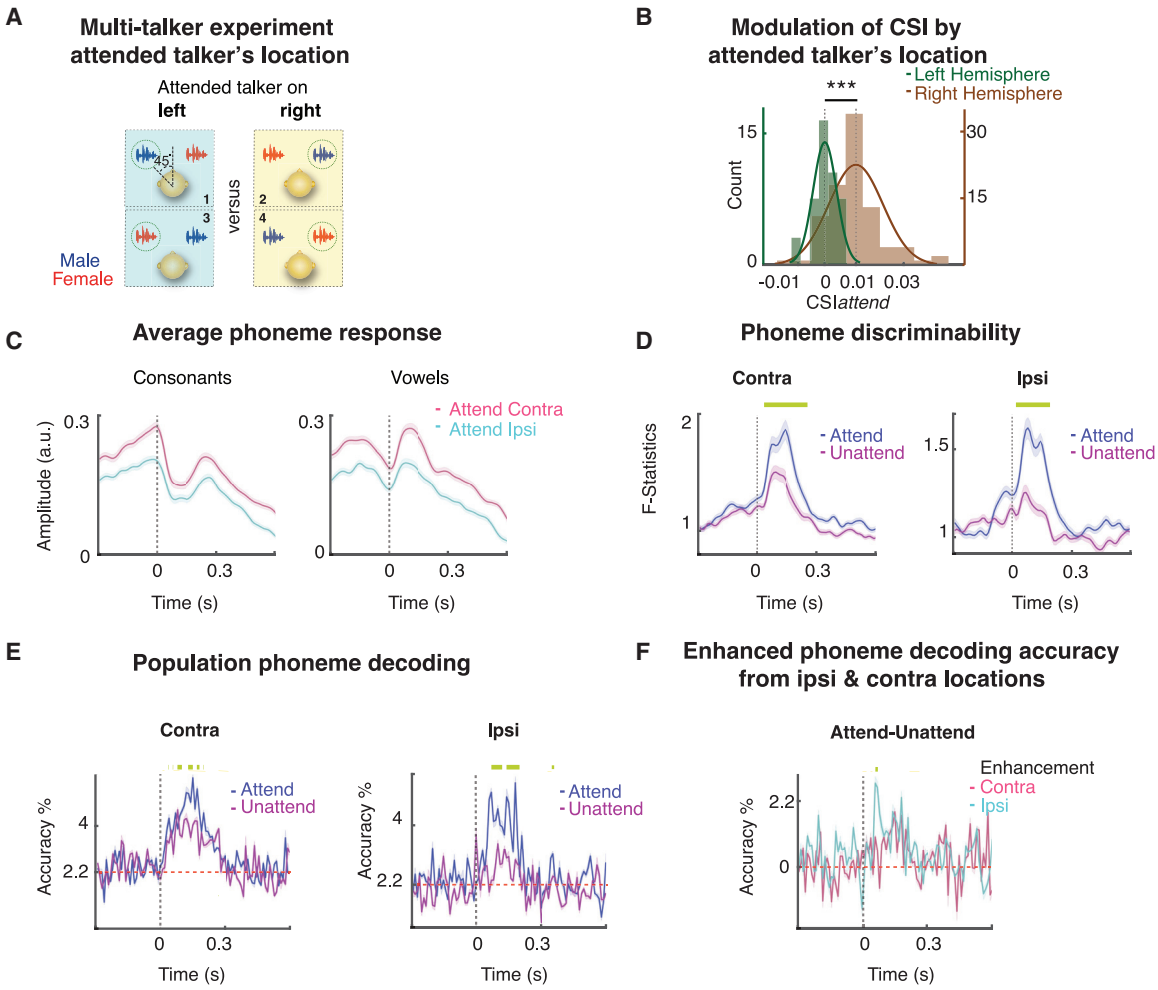

Figure 4. Attended talker's location modulates the neural MRL

(A) Task schematic. Spatially separated multi-talker experiment. Comparing neural responses for attended talker's location left versus right conditions (blue versus yellow blocks).

(B) Histograms of contralateral selectivity index for the attended talker (CSI*attend*) from mean high gamma responses of individual electrodes for the left and the right brain hemispheres ($p < 0.001$, unpaired t test). See also Figure S2.

(C) Average high gamma responses to phoneme consonants (left) and vowels (right) from the two attended talker's locations, contralateral in pink and ipsilateral in blue.

(D) Average F-statistic from single electrodes for the separation of 46 individual phonemes for attended (blue) and unattended (magenta) conditions ($p < 0.001$, paired t test). The plot for the contralateral location is on the left, and the plot for the ipsilateral location is on the right.

(E) Population phoneme decoding accuracy (y axis) for 46 individual phonemes from attended (blue) and unattended (magenta) conditions ($p < 0.001$, paired t test). The plot for the contralateral location is on the left, and the plot for the ipsilateral location is on the right.

(F) Enhancement of phoneme decoding accuracy with the location of attended talker from contralateral (pink) and ipsilateral (blue) locations ($p < 0.001$, paired t test).

the electrodes had significantly higher STRF predictions for contralateral speech than for ipsilateral speech ($p < 0.001$, permutation test) (Figure 2B). We quantified this preference for contralateral talkers by defining a metric called spatial tuning (ST). ST is defined as the difference in STRF prediction accuracy (Pearson's r) of speech between contralateral and ipsilateral locations normalized by the sum of the two (STAR Methods). We found that neural sites had higher ST in the spatially separated multi-talker scenario than in the single-talker case ($p < 0.001$, paired t test; Figure S1B), indicating a more accurate encoding of contralateral talkers compared with that of ipsilateral talkers in the multi-talker scenario. However, as averaging over attended and unattended conditions may obscure weak effects in one condition when there is a particularly strong effect in the other condition (e.g., weak bottom-up effects in the unattended condition may be obscured by stronger top-down attentional effects in the attended condition), we performed a follow-up control analysis. That is, to ensure that the observed contralateral preference is independent of attention, we repeated the analysis from Figure 2B separately for both the attended condition and unattended condition. As expected, we found that electrodes had higher STRF prediction correlations for attended

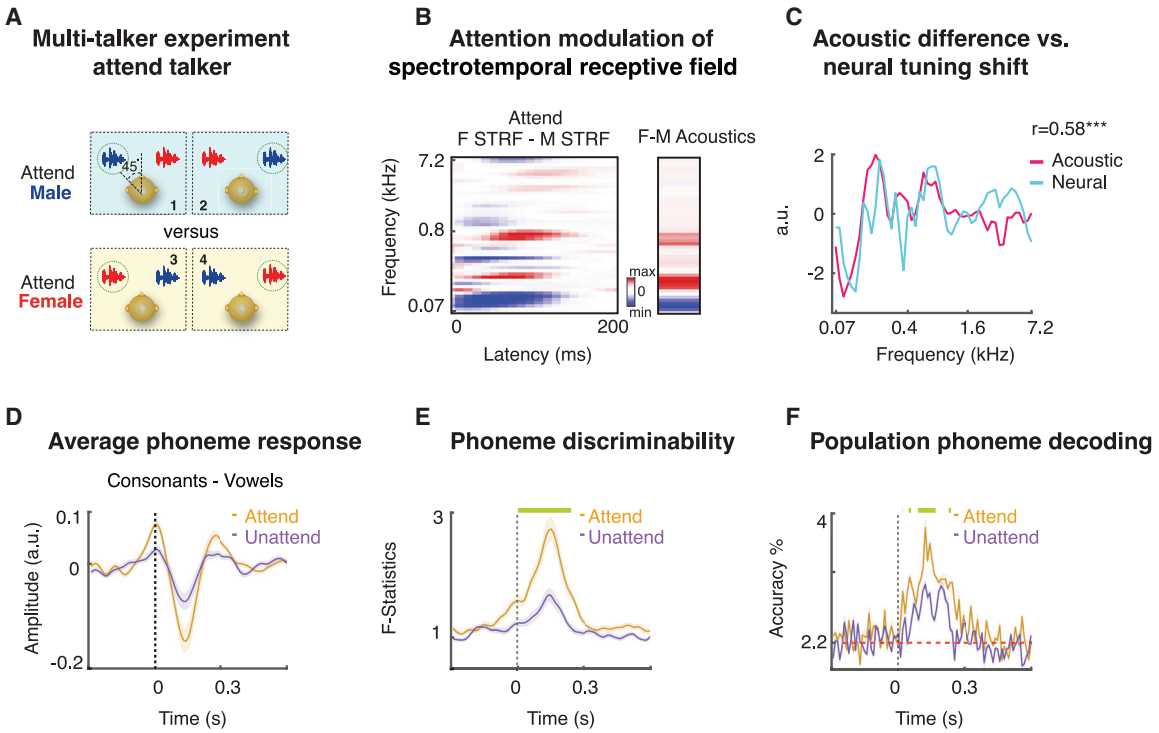

Figure 5. Attention to the identity of the talker modulates the neural STRF

(A) Task schematic. Spatially separated multi-talker experiment. Comparing neural responses in the attended male and attended female conditions (blue versus yellow blocks).

(B) Average STRF from all AMI electrodes for the attended female minus the attended male condition for the same mixed stimulus (left); attended female minus attended male acoustic profile from single-talker speech spectrograms (right). See also Figures S3 and S4.

(C) First principal component of the neural STRF from (B) in blue; acoustic spectral profile in pink.

(D) Average high gamma responses to consonants minus vowels for the attended talker condition in orange and the unattended condition in purple.

(E) F-statistic from single electrodes for the separation of 46 individual phonemes from the attended talker (orange) and unattended talker (purple) conditions ($p < 0.001$, paired t test).

(F) Population phoneme decoding accuracy (y axis) of 46 individual phonemes from attended speech (orange) and unattended speech (purple) ($p < 0.001$, paired t test).

contralateral speech compared with attended ipsilateral speech (Figure S1D, $p < 0.001$, paired t test) and higher STRF prediction correlations for unattended contralateral speech compared with unattended ipsilateral speech (Figure S1E, $p < 0.001$, paired t test). These results confirm that the contralateral preference in the multi-talker condition is independent of top-down attentional effects.

To examine the encoding of phonetic features in the multi-talker case, we measured the difference between average responses to consonants and vowels. In contrast to single-talker scenarios (Figure 1F), we found that in the multi-talker scenario, the qualitative response to phoneme features has a larger dip when computed for the talker in the contralateral location (Figure 2C, pink) than for the talker in the ipsilateral location (Figure 2C, blue). We quantified this effect for all individual electrodes by measuring phoneme discriminability (F-statistics) for contralateral and ipsilateral talkers using 46 individual phonemes. We found that the F-statistic was indeed higher for contralateral speech than for ipsilateral speech (Figure 2D, $p < 0.001$, paired t test). Population decoding of 46 individual phonemes, using a linear classifier, also shows superior decoding accuracy for phonemes of the contralateral talker (Figure 2E, $p < 0.001$, paired t test). Together, these results show that in a multi-talker condition with spatially separated talkers, the neural sites preferentially represent the talker on the contralateral side, as compared with the talker on the ipsilateral side, and more accurately encode the spectrotemporal and phonetic features of the contralateral talker. This enhanced response to the contralateral talker is a bottom-up effect, as it occurs irrespective of top-down attention. Note that we also evaluated whether this effect was stronger in primary (HG) or higher-order (PT, STG) auditory regions, but we did not find a significant difference between the ST from different anatomical regions in the auditory cortex (HG, PT, and STG; Figure S1C, $p > 0.05$, unpaired t test). To better visualize ST from these anatomical areas we have color-coded Figures 1D and 2B by anatomical region (Figures S1F and S1G).

To investigate whether the observed ST shift to the contralateral location is caused by the presence of an additional competing talker or whether it depends on spatial separation between the talkers, we designed a control condition in which the two simultaneous talkers were collocated in space (at either 45 or −45 degrees; Figure 3A). Due to the limitations in recording time with patients, we could run this control in only two out of

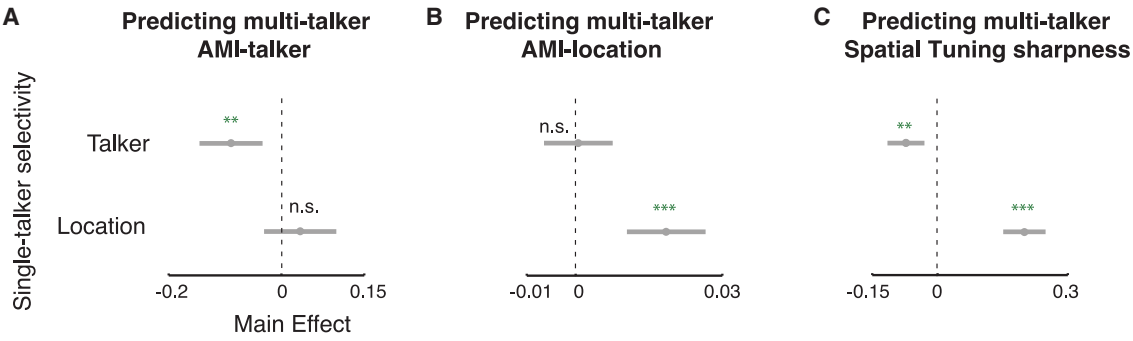

Figure 6. Relationship between single-talker neural features and multi-talker neural features

(A) Top-down attention modulation to the talker identity in the multi-talker case is explained by single-talker selectivity but not by location selectivity. See also Figure S4.

(B) Top-down attention modulation to talker location in the multi-talker case is explained by single-talker location selectivity but not by talker selectivity.

(C) Bottom-up sharpening of neural tuning to contralateral talker in the multi-talker case is explained by single-talker location selectivity and by talker selectivity.

seven subjects (40 out of 119 responsive electrodes). We repeated the analysis from Figure 2B for the collocated multi-talker scenarios. In contrast to the spatially separated multi-talker condition (Figure 2B), in the collocated multi-talker condition we found the lack of ST to speech from the contralateral location compared with the ipsilateral location (Figure 3B, $p < 0.05$, paired t test). However, similar to Figure 2B, neural responses for spatially separated multi-talkers for these two subjects show preference in encoding contralateral speech over ipsilateral speech (Figure 3C, $p < 0.001$, paired t test). We also verified that the control analysis that we performed in the spatially separated multi-talker condition (i.e., in which we compute the contralateral preference separately for the attended and unattended condition; Figures S1D and S1E) holds true when performed on these two subjects (paired t test, $p < 0.001$). We further repeated the phoneme discriminability analysis from the spatially separated multi-talker scenario (Figures 2D and 2E) for the spatially collocated multi-talker scenario. In contrast to the spatially separated multi-talker scenario, we found that the phoneme discriminability (F-statistics) for individual electrodes in the spatially collocated scenario was not higher when the talkers were on the contralateral side of electrodes than on the ipsilateral side (Figure 3D, $p > 0.05$, paired t test). Similarly, the population decoding analysis did not show higher accuracy in decoding the phoneme features of the talkers on the contralateral side (Figure 3E, $p > 0.05$, paired t test). In summary, we found that for spatially separated talkers, the electrodes in the auditory cortex become inherently more responsive to the speech of the contralateral talker than that of the ipsilateral talker and better represent the spectrotemporal and phonetic features of the contralateral talker's speech. However, this contralateral benefit only occurs when there is spatial separation between talkers. When talkers are collocated, there is no effect of location on speech encoding similar to our observations in the single-talker scenario.

Top-down modulation of neural responses by the location of the attended talker

We showed that in a single-talker experiment, the location of the talker changes the MRL of the neural responses, where on average, the mean is higher for contralateral locations. In this section, we examine how the location of an attended talker in a spatially separated multi-talker condition changes the neural response. Furthermore, we perform additional analysis to examine the lateralization of the alpha band based on the attended talker's location. Finally, we examine the anatomical distribution of brain areas that show varying neural representation based on the attended talker's location. We addressed this question by combining the neural response of each electrode to all trials in which the attended talker was at the contralateral location and compared it with all trials in which the attended talker was at the ipsilateral location (Figure 4A, blue versus yellow), irrespective of the talker's voice (i.e., male or female). In this comparison, the same speech tokens were heard in each condition, the only difference was that the attended talker was at left versus right locations. We quantified attentional effects by calculating the MRL and the resulting CSI for each electrode in the *attend*-left and *attend*-right conditions (STAR Methods), mirroring the analysis for the single-talker scenarios (Figure 1C). The measure "CSI*attend*" for all electrodes is shown in Figure 4B, and histograms are colored by the electrode location in the left (green) or right (brown) hemisphere. Figure 4B shows that CSI*attend* is positive for the right hemisphere electrodes, indicating their higher MRL in the attend-left (contralateral) condition compared with the attend-right (ipsilateral) condition. However, the CSI*attend* for left hemisphere electrodes is centered around zero. The difference between means of the histograms is 0.012 (Figure 4B, $p < 0.001$, unpaired t test). Note that this difference is smaller than the difference obtained in the single-talker case (0.03), showing the weakening of ST in the presence of competing talkers. This analysis thus shows that the electrode's MRL is modulated by the attended talker's location, where the contralateral location increases the relative neural MRL, compared with the ipsilateral location. The similarity of CSI*attend* in the multi-talker condition and CSI in the single-talker condition shows a high correlation between the two values ($r = 0.53$, $p < 0.001$; Figure S2A). This high correlation shows that the more selective an electrode is to the location of speech in the single-talker case, the more will it be modulated by attention to location in the multi-talker case. The CSI*attend* is for each electrode is shown on the brain in Figure S2B. We did not find a significant difference between

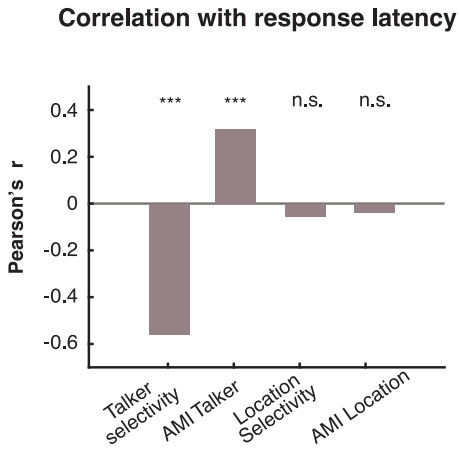

Figure 7. Relationship with the latency

Correlation values between electrode response latency and the effects from single-talker and multi-talker conditions.

See also Figure S5.

the absolute value of CSI*attend* in different anatomical regions (HG, STG, and PT) in the auditory cortex (Figure S2C, $p > 0.05$, unpaired t test). Replicating the analysis in Figure 4B separately for each talker also reveals similar trends in CSI (Figure S2D). Finally, we analyzed attentional modulations of power in the alpha band to verify if our dataset reflects the well-known alpha lateralization[34] in response to directional attention and found similar contralateral lateralization as reported by previous studies (Figure S2E).

Next, we examined how the attended talker's location changed the encoding of phoneme features of the talker arriving from that location. The average response to consonants (Figure 4C, left) and vowels (Figure 4C, right) shows that the qualitative baseline response to phoneme features is higher for attended contralateral speech (in pink) than for attended ipsilateral speech (in blue) (Figure 4C). The raw average phoneme plots qualitatively show that the location of the attended talker indeed modulates the MRL of neural sites. Furthermore, we quantified whether the location of an attended talker enhanced the phoneme discriminability (F-statistics) of speech arriving from that location at the level of individual electrodes. Phoneme discriminability of 46 individual phonemes for the attended location is shown in Figure 4D in blue, which is significantly higher than when the same speech is uttered from the unattended location in magenta ($p < 0.001$, paired t test). Similarly, phoneme decoding from the population of electrodes revealed that the attended talker's location enhances the decoding accuracy of individual phonemes in that location (Figure 4E, $p < 0.001$, paired t test). Interestingly, attention has a larger effect in the ipsilateral location than in the contralateral location (Figure 4F, $p < 0.001$, paired t test), presumably because the contralateral location is already enhanced due to bottom-up effects, as shown in Figure 2.

In summary, we found that the location of the attended talker modulates the MRL of the neural responses with higher MRL when an attended talker is on the contralateral location, and this modulation enhances the encoding of phoneme features at the location of the attended talker.

Top-down modulation of neural responses with attention to the talker's voice

The previous analysis shows how an attended talker's location modulates the MRL of the neural response. In this section, we examine how the attention to a talker's voice irrespective of location modulates the neural response. We also perform additional analysis to explore the effect of this modulation in different anatomical regions (HG, PT, and STG) and in right versus left brain hemispheres. Further, we qualitatively show these effects in single electrodes with raw data plots.

To answer the central question of the effects of attention on neural representation of a talker's voice, we combined the neural response of each electrode to all trials when the subject attended one talker and compared the response with all conditions when the subject attended the other talker, irrespective of where the talker was located (Figure 5A, comparing conditions blue versus yellow). We evaluated the effect of attention to the talker by measuring the spectrotemporal tuning of neural sites in a spatially separated multi-talker scenario. The STRF of electrodes are estimated from the same mixed stimuli for two conditions: first in the attend-male condition and second in the attend-female condition (STAR Methods; Figure 5A). Since the analysis is performed on the same mixed audio, the differences in STRFs directly reflect the effects of attention to talker voices on the STRFs. We have previously shown that attentional effects are not equally present in all neural sites.[3] Hence, we first defined an attentional modulation index (AMI), designated AMI-talker, as the difference in STRF predictions from attended and unattended talkers across both locations normalized by their sum (Figure S3A).[3,4] We found that 44 neural sites (36.97%) were significantly modulated by attention to the talker (AMI-talker > 0.04, threshold chosen based on null distribution of AMI-talker) (STAR Methods).[3] For these electrodes, we measured two STRFs, one that mapped the mixed stimulus spectrogram to the attend-male neural response and another that mapped the same mixed stimulus to the attend-female neural response. We then averaged these STRFs across all AMI-modulated electrodes and calculated the difference between the resultant average attend-female STRF and average attend-male STRF (Figure 5B, left). The colors red and blue in Figure 5B reflect the difference between excitatory regions in average attend-female and average attend-male STRFs. The difference in STRF shows the change in frequency tuning with positive values (red) indicating increased responsiveness to frequencies when attending to the female talker and negative values (blue) indicating increased responsiveness to frequencies when attending to the male talker. We further compared this difference in STRF with the difference between acoustic spectral power of female and male talkers (Figure 5B, right), obtained by measuring the difference between time-averaged spectrograms of the female and male talkers, respectively (STAR Methods). The difference in spectral power highlights the contrast between their pitch and formant frequencies. The similarity between the neural STRF difference and the acoustic spectral power difference in Figure 5B showed that attention to a talker shifts the spectral tuning of electrodes toward the distinctive spectral features of that talker. For example, the higher formant frequency of the female talker (in red) or the lower pitch of the male talker (in blue) in Figure 5B. We quantified this

talker-matched tuning shift by comparing the first principal component of the STRF in Figure 5B with the spectral difference of talkers (Figure 5C; $r = 0.58$, $p < 0.001$). The colormap in Figure 5B was adjusted for visualization but the effect is quantified without adjusted STRFs in Figure 5C. The tuning shift in the STRF also holds true when each talker location was considered separately (Figures S3B, S3C, S3E, and S3F). Note that we used single-channel STRFs (i.e., STRFs measured using nonspatial, single-channel audio) for our analysis as previous studies have shown independent encoding of spatial and spectrotemporal features in the human auditory cortex.[12] However, to verify an unbiased STRF measurement that includes spatial features, the analysis was repeated and verified by concatenating the left and right stereo STRF (Figures S3G and S3H). Further, the STRF analysis was performed separately for left and right brain hemispheres in Figure S3I to verify that the same modulation trends persist in each brain hemisphere. We did not find significant differences between the absolute value of AMI from different anatomical regions in the auditory cortex (Figure S3D, $p > 0.05$, unpaired t test).

To examine the consequence of this tuning shift on phoneme encoding of the attended talker, we compared the phoneme discriminability of attended and unattended talkers. Figure 5D shows the averaged neural response to consonants minus vowels for attended (in orange) and unattended talkers (in purple) where a larger response to the phoneme features of the attended talker can be seen. Phoneme discriminability (the F-statistics) also shows significantly better separation for the individual phonemes of the attended talker (Figure 5E, $p < 0.001$, paired t test). Finally, the population neural decoding of individual phonemes also shows superior accuracy for the attended talker (Figure 5F, $p < 0.001$, paired t test). We did not find a relationship between modulation due to attention to location and attention to talker ($r = -0.2$, $p > 0.05$; Figure S4A). In sum, we found that attention to a specific voice changes the spectrotemporal tuning of neural sites to match the distinctive acoustic features of that talker. The result of this shift is the enhancement of phoneme encoding of the attended talker relative to the unattended talker.

Relationship between neural responses in single-talker and multi-talker conditions

We showed that in the single-talker condition, the location of the talker changes the mean neural response. In multi-talker experiments, we observed three effects: (1) stimulus-driven enhancement of responsiveness to speech from the contralateral location; (2) modulation of MRL with the location of the attended talker; and (3) modulation of spectrotemporal tuning with attention to the talker's voice. Here, we examined how the response properties in single-talker and multi-talker experiments are related across individual electrodes.

From the single-talker experiment, we measured two parameters for each electrode, quantifying the selectivity of electrode responses for a specific talker and for a specific location. Talker selectivity is the difference in the neural response to the two talkers, and it indicates the degree of preferred response of an electrode to one of the two talkers as quantified by a two-sample t test (STAR Methods).[3] Location selectivity is the absolute value of CSI (Figure 1; STAR Methods) and reflects the degree of change in the mean neural response with respect to the location. From the multi-talker experiment, we quantified three parameters for each electrode, quantifying the degree of spatial tuning sharpness to contralateral talker (ST), attentional modulation to attended talker's location (AMI-location), and attentional modulation to talker (AMI-talker) (STAR Methods).

To find the relationship between single-talker and multi-talker parameters, we performed linear regression to predict the multi-talker parameters (ST, AMI-location, AMI-talker) from the single-talker parameters (location and talker selectivity) (STAR Methods). First, we found that talker selectivity successfully predicted the attentional modulation to talker and did so with negative weight, meaning that the electrodes whose response was more similar between the two talkers (or in other words not preferentially "tuned" to one of the talkers) showed higher attentional modulation when attending to a talker in multi-talker conditions (Figure 6A). On the other hand, the electrodes that was significantly more responsive to a specific talker (prefers a talker due to its spectrotemporal tuning) changed less with attention. Example electrode responses are provided in Figure S4C, showing how a high talker selective electrode encodes phonemes of preferred talker regardless of attentional focus, and how a high attention modulated electrode encodes phonemes of an attended talker irrespective of talker identity (Figure S4C). Single-trial examples for AM are also shown in Figure S4E. Second, we found that location selectivity predicts both ST and AMI-location but not AMI-talker (Figures 6B and 6C). This result meant that electrodes showing a preferred location in the single-talker experiment also exhibited a larger shift to the contralateral talker (bottom-up) and were also more highly modulated by the attended talker's location (top-down). However, we did not find a direct correlation between ST and AMI-location ($r = 0.04$, n.s.; Figure S4B), suggesting that location selectivity in the single-talker case governed ST and AMI-location in the multi-talker case through two disjointed mechanisms. ST exhibited by single electrodes is demonstrated in Figure S4D. Last, we found that while both talker and location selectivity are significant predictors of the ST, location selectivity has a higher contribution to the prediction of ST than talker selectivity. Overall, this analysis establishes a direct connection between the bottom-up stimulus-encoding properties of an electrode and how these properties change with top-down attention, therefore shedding light on the interaction of these two mechanisms in spatial multi-talker conditions.

Latency of attentional modulation to location and talker

Our multi-talker results show that the location of attended talker (AMI-location) and the voice of attended talker (AMI-talker) differentially and independently modifies different aspects of neural responses. We did not observe clear anatomical and topographical organization of attentional modulations (Figures S2C and S3D). However, anatomical divisions may not fully reflect the underlying functional organization of auditory cortical responses.[45] In addition, our intracranial recordings only sparsely sample the anatomical regions, which makes functional analysis challenging. Here, we used the latency of electrode responses to approximate how high the electrodes are in the neural auditory processing hierarchy, as the latency reflects the number of synapses away from the auditory periphery.[46] We then looked to see if the latency of electrodes was

related to the attentional processes. We measured the response latency as the timing of the excitatory peak of the electrode's STRF (STAR Methods). In Figure 7, we find that the electrode's latency is significantly correlated with both talker selectivity (negatively; $r = -0.56$, $p < 0.001$) and AMI-talker (positively; $r = 0.31$, $p < 0.001$) (see also Figures S5A, S5B, and S5E), indicating that the sites in higher auditory areas lose talker selectivity but become more strongly modulated by attention to a talker. The combination of increased attentional modulation of electrodes with lower single-talker selectivity and their higher latency suggests that higher levels of neural processing are less selective to specific acoustic features that enable stronger plasticity to better represent the attended talker. This finding is in line with a previous intracranial study showing selective representation of talker's voice in primary auditory cortex that is not modulated by attention and representation of the attended talker in the higher nonprimary auditory cortex.[3] However, the lack of correlation between latency and locational selectivity ($r = -0.06$, $p > 0.05$), on the one hand, and the lack of correlation between latency and AMI-location ($r = -0.04$, $p > 0.05$), on the other, shows that these locational effects do not emerge progressively and are scattered throughout the auditory cortex, possibly reflecting their subcortical origin[47] (Figure 7; see also Figures S5C and S5D).

Together, these findings suggest that the modulatory effects of the attended talker's location start earlier in the auditory pathway and do not increase through the neural processing hierarchy. The modulatory effects of attention to the talker's voice, however, are progressive and emerge later. Since attention to the talker happens later in the pathway, this process possibly requires more complex neural computations compared with those for locational attention.

DISCUSSION

We use direct neural recordings from the human auditory cortex to examine the bottom-up and top-down effects that shape the neural representation of speech in single-talker and multi-talker acoustic conditions where talkers are either spatially separated or spatially overlapping. In the single-talker condition, we find that whereas the location of the talker changes the MRL, the encoding accuracy of spectrotemporal and phoneme features is not modulated by location. By contrast, the neural responses in the spatially separated multi-talker condition encode the spectrotemporal and phonemic features of the contralateral talker compared with the ipsilateral talker with higher fidelity, irrespective of the subject's attentional focus. However, attention differentially modulates either the mean response or the spectrotemporal tuning of neural sites, depending on whether the modulation is to the attended talker's location or voice, respectively. These attentional modulations help with the decoding of the attended speech. We observed the effect of attentional modulation to the talker's voice only in the higher auditory cortex responses, but attentional modulation due to the talker's location is present throughout the auditory cortex. Finally, we found that the spatial and spectrotemporal tuning properties of neural sites in single-talker conditions accurately predict the degree and type of tuning change that would be mediated by top-down attention in a multi-talker condition, suggesting a specialized role for various neural populations in representing multi-talker speech.

Stimulus-driven sharpening of spatial tuning

We found a stimulus-dependent and attention-independent increased preference for encoding of the contralateral talker in the spatially separated multi-talker condition. A pre-attentive preference to encode the contralateral sound source has also been shown in passive anesthetized animals,[19,20] which further emphasizes the bottom-up nature of this tuning shift. Previous studies have further shown that this bottom-up tuning shift results in separation of the competing sources from different locations in the brain, as each location predominantly activates a different population of neurons, resulting in separable encoding of sound sources from different locations.[20] This finding is consistent with our result, which also shows better phoneme feature encoding for the contralateral talker over the ipsilateral talker in the spatially separated multi-talker scenario. Our study extends the findings from animal models in several ways. First, we show that such a bottom-up pre-attentive separation of the content of simultaneous spatially separated sounds also occurs for a complex natural sound such as speech and in awake, behaving humans. Second, we show that the effect of ST is not present when there is no spatial separation between simultaneous sound sources, in which case such a shift cannot facilitate stream segregation. Last, we show that this sharpened ST still remains plastic and can be modulated with top-down attention. Since we find these processes irrespective of attentional focus, it suggests that they may be the result of anatomically hard-wired neural circuits such as the neural circuitry that enables lateral inhibition in the auditory periphery. Such a bottom-up representation thus gives a substrate on which top-down auditory attention may act by enhancing or suppressing the behaviorally relevant streams. The animal literature remains unclear as to whether sound localization and stream segregation from spatial cues share a common anatomical pathway[19,48] or whether they are disassociated.[49] In our study, we find 61.7% of the electrodes that are modulated by the attended talker's location also exhibit bottom-up sharpening of ST, which suggests a shared neural substrate for these two effects. However, the degree of shift caused by bottom-up ST to the location and the top-down modulation of responses with the location of the attended talker was not correlated. This lack of correlation implies the possibility that these two effects are governed by separate neural mechanisms, where one represents the stimulus features and the other integrates top-down signals into this representation. It is worth noting that the bottom-up contralateral preference we report here has not always been seen in human noninvasive studies,[33,34,50] which could be due to the lack of resolution necessary to capture such cortical effects and disentangle them from attention-driven effects.

Modulation of neural responses by the attended talker's location

The modulation of neural responses by the attended talker's location in a spatial multi-talker scenario has been studied in humans only with noninvasive methods. EEG and magnetoencephalography (MEG) studies report the contralateral lateralization of the alpha band (8–12 Hz) in response to the attention to

location of a talker.[33,34] The underlying neural mechanisms that lead to alpha band lateralization can only be speculated due to the resolution limitations of neural measurement methods. Here, we show that the location of an attended talker in the multi-talker condition changes the MRL of the neural response in the high gamma band. This modulation is stronger in the right brain hemisphere than in the left brain hemisphere. This finding is consistent with an EEG study that also showed stronger attentional modulations in the left hemisphere.[51] However, it should be noted that electrode coverage is limited in the present study. Further studies with comparable hemispheric coverage are necessary to verify this lateralization. Since it has previously been shown that the high gamma band is related to the firing rate of the underlying population of neurons close to the electrode[36,52], a change in MRL with attention to location is likely a change in the baseline firing rate of the underlying neuronal population that encodes that location. Whether the lateralization in the alpha and high gamma bands is mediated by the same correlative or causal neural mechanism requires further research.

Extensive work has been done in the cognitive science literature to pin down the top-down attention control mechanisms,[53,54] including delineating the involvement of right temporoparietal junction (RTPJ) and the left inferior parietal supramarginal part (LIPSP) top-down attention control based on attention switching, using spatial and pitch features, respectively;[1] and the involvement of dorsal pre-central sulcus, superior parietal lobule, and left frontal eye fields (FEF) versus inferior frontal gyrus and left posterior STS for the preparatory activity, using spatial attention versus pitch attention, respectively.[2,55] Our study is focused on modulation of neural response in the auditory cortex based on attended spatial features as opposed to attention control mechanisms in the non-auditory areas. Further research can possibly shed light integrating how top-down attention control and top-down attention selection mechanisms work together to separate the attended speech.

Modulation of neural responses by attention to the talker's voice

Neural modulation by attention to a talker's voice has also been extensively studied in humans with invasive[3,4] as well as noninvasive methods.[32,56,57] These studies find a static and diverse tuning to spectrotemporal features of talkers in primary auditory areas and increased dynamic modulation by attention to talkers in the nonprimary auditory areas. Our results are consistent with this finding, as we also find increased attentional modulation effects to a talker (AMI-talker) in areas with longer response latency. However, we did not see the same trend in the degree of attentional modulation to location (AMI-location), as it did not show a correlation with latency. One implication for this finding is that top-down attention is integrated into the representation only at a level that has the required representational complexity to separate sound sources in the bottom-up hierarchy. This notion therefore suggests an earlier and possibly subcortical origin for processing spatial information, which is plausible as hemispheric interactions start early in the auditory pathway before the signal reaches the cortex.[47] Talker separation, on the other hand, is different. Because of the acoustic similarity of speech from different talkers, their representation is highly overlapping in the auditory periphery, as they share similar spectral and temporal energies. Therefore, talker separation first requires the transformation of the time-frequency representation of the acoustic signal at the auditory periphery onto a high-dimensional representation in which the talkers are more easily separable,[3,4] which can be achieved using nonlinear transformations of the acoustic signal.[58] Our findings are in line with more recent studies showing transformation of neural responses from acoustic to phonetic to lexical information with an increase in latency[50] and differential modulation of acoustic versus higher-order phonetic and lexical representation by attention.[50,59]

The formation of auditory objects in complex auditory scenes depends on poorly understood interactions between goal-driven top-down and stimulus-driven bottom-up processes. The goal-driven top-down attentional process selectively allocates cortical processing resources to sounds of interest, whereas the stimulus-driven bottom-up process registers the entire acoustic scene and selectively gates incoming salient signals.[60] Past noninvasive human studies have shed light on the anatomical and functional categorization of these two processes. These studies show that largely overlapping neural circuits are activated by bottom-up and top-down auditory attentional processes.[61,62] The top-down effects are said to enhance the neural encoding of voluntarily attended sensory inputs relative to other objects in the scene and suppress irrelevant sensory information.[63] Similarly to our result, attention is shown to modulate different aspects of the neural response, depending on where the attention is directed. For example, depending on the task and the focus of attention, neurons can shift their spectral tuning,[27] temporal modulation tuning,[64] spectral and temporal structure of the stimulus,[65] and ST during active sound localization.[25,26] Our findings are consistent, as we show that attention to the location or voice of a talker has differential effects on neural responses in multi-talker speech. From the encoding perspective, previous studies show a bottom-up multidimensional and multiplexed representation of sound in the mammalian auditory cortex.[18,66,67] These studies show that the same neural population in the primary auditory cortex encodes stimulus features along both spectral and spatial features. Consistent with these findings, we also found that spectral and spatial information are encoded in different aspects of the neural response and in the same population. Our study, however, goes beyond this finding by also showing that the neural encoding of spatial and spectral features is independently modulated by top-down attention to these stimulus properties. Collectively, these results demonstrate a neural coding scheme in which different aspects of the stimulus are separately encoded by bottom-up mechanisms that then interact with top-down attention to these features, therefore creating a multifeatured, multidimensional, and dynamic representation of the stimulus.

Contribution to the hypothesized sound localization frameworks

The traditional theoretical framework has postulated a dual hierarchical "what" and "where" pathway[68] with an anteroventral processing stream responsible for the recognition of auditory objects such as communication sounds and a posterodorsal stream responsible for sound localization. With this view, the

cortical processing of sound location is expected to be found in the higher posterior-dorsal auditory areas, such as the planum temporale. However, the extent to which the neural mechanisms involved in sound localization align with this model is a matter of debate.[30,69] On the one hand, there is evidence in support of such a hierarchical processing of sound location in both humans[70,71] and animals,[72,73] showing that the location of sound specifically affects higher-order areas posterior to the primary auditory cortex. On the other hand, few recent studies in animals[25] and humans[12,26] show that the location of sound stimulus affects neural responses as early as the primary auditory cortex, going against a strict hierarchical feedforward model of spatial processing.[30] Our study provides insight by identifying the neural populations that are affected by these effects. Looking at the neural processing hierarchy in the auditory pathway, we find that (1) the bottom-up ST, (2) the top-down AMI-location, and (3) the top-down AMI-talker target different but overlapping cortical neural populations. We find that the effects of bottom-up sharpening of neural ST and top-down modulation by attended location start as early as the primary auditory cortex, and top-down attention to talker identity is largely dominant in neural populations later in the auditory pathway. The former suggests that location processing is not strictly feedforward since we find effects of attentional modulation on location in short latency electrodes in the primary auditory cortex and that this processing does not increase throughout the processing hierarchy. Our findings are thus in line with the recent work that active, goal-oriented localization reduces the assumptions of increased spatial processing in the posterodorsal stream, which was often established using passive listening.[25,26]

An important debate concerning sound location processing is whether one brain hemisphere is sufficient to localize sound or whether both hemispheres are necessary.[30] Decoding studies report better location estimates when both brain hemispheres are employed.[12,17] We report that bottom-up sharpening of ST enhances the encoding of contralateral speech, and top-down attention to location changes the baseline MRL of these same sites. In particular, we find that attending to the ipsilateral side has a larger modulatory effect than attending to the contralateral side, presumably because the contralateral talker is already enhanced by bottom-up ST. Hence, one possible contribution from our results to this theoretical debate is a model of spatial processing in which bottom-up ST separates sources to different brain hemispheres. Top-down attention to the location acts as a rectifying network indicating which of the two brain hemispheres should be given a larger weight in order to readout the attended talker's speech from the neural response. Further studies using connectivity analysis can shed more light on this question by determining how network connectivity between hemispheres changes when a subject switches attention from one location in space to another.

By examining the bottom-up representational properties and top-down attentional modulation of spatially separated multi-talker speech in the human auditory cortex, our study takes a major step toward determining the neural mechanisms that are involved in processing spatially separated multi-talker speech and the interaction and integration of bottom-up and top-down signals giving rise to signal transformations that allow for the selection of a desired stream from the multitude of sound sources.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Participants and data collection
- METHOD DETAILS
  - Stimulus
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Preprocessing neural data
  - Calculating the spectrotemporal receptive fields (STRF)
  - Brain maps
  - Estimation of the metrics of CSI, CSI*attend*, location selectivity, AMI-location, talker selectivity, AMI-talker, and ST
  - Phoneme F-Statistics and decoding analysis
  - Calculating the acoustic difference between talkers
  - Linear regression analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.07.036.

ACKNOWLEDGMENTS

We thank Richard Thompson Lee for their comments on the manuscript text. This work was supported by National Institutes of Health grant R01DC018805 and National Institute on Deafness and Other Communication Disorders grant R01DC014279.

AUTHOR CONTRIBUTIONS

P.P. and N.M. designed the experiment. P.P., N.M., J.L.H., S.B., and A.D.M. recorded the data. P.P. and N.M. analyzed the data. P.P. and N.M. wrote the original manuscript. K.v.d.H., P.P., and N.M. edited the manuscript. All authors commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 11, 2022
Revised: June 8, 2022
Accepted: July 19, 2022
Published: August 15, 2022

REFERENCES

1. Larson, E., and Lee, A.K.C. (2014). Switching auditory attention using spatial and non-spatial features recruits different cortical networks. Neuroimage *84*, 681–687.
2. Lee, A.K.C., Rajaram, S., Xia, J., Bharadwaj, H., Larson, E., Hämäläinen, M.S., and Shinn-Cunningham, B.G. (2012). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. Front. Neurosci. *6*, 190.

3. O'Sullivan, J., Herrero, J., Smith, E., et al. (2019). Hierarchical encoding of attended auditory objects in multi-talker speech perception. Neuron *104*, 1195–1209.e3.
4. Mesgarani, N., and Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature *485*, 233–236.
5. Macken, W.J., Tremblay, S., Houghton, R.J., Nicholls, A.P., and Jones, D.M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. J. Exp. Psychol. Hum. Percept. Perform. *29*, 43–51.
6. Sussman, E.S., Horváth, J., Winkler, I., and Orr, M. (2007). The role of attention in the formation of auditory streams. Percept. Psychophys. *69*, 136–152.
7. Shinn-Cunningham, B.G. (2008). Object-based auditory and visual attention. Trends Cogn. Sci. *12*, 182–186.
8. Fritz, J.B., Elhilali, M., David, S.V., and Shamma, S.A. (2007). Auditory attention—focusing the searchlight on sound. Curr. Opin. Neurobiol. *17*, 437–455.
9. Shamma, S.A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. Trends Neurosci. *34*, 114–123.
10. Cusack, R., Deeks, J., Aikman, G., and Carlyon, R.P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J. Exp. Psychol. Hum. Percept. Perform. *30*, 643–656.
11. Niebur, E., Hsiao, S.S., and Johnson, K.O. (2002). Synchrony: a neuronal mechanism for attentional selection? Curr. Opin. Neurobiol. *12*, 190–194.
12. 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. https://doi.org/10.1016/j.celrep.2018.07.076.
13. Ortiz-Rios, M., Azevedo, F.A.C., Kuśmierek, P., Balla, D.Z., Munk, M.H., Keliris, G.A., Logothetis, N.K., and Rauschecker, J.P. (2017). Widespread and opponent fMRI signals represent sound location in macaque auditory cortex. Neuron *93*, 971–983.e4.
14. Woods, T.M., Lopez, S.E., Long, J.H., Rahman, J.E., and Recanzone, G.H. (2006). Effects of stimulus azimuth and intensity on the single-neuron activity in the auditory cortex of the alert macaque monkey. J. Neurophysiol. *96*, 3323–3337.
15. Harrington, I.A., Stecker, G.C., Macpherson, E.A., and Middlebrooks, J.C. (2008). Spatial sensitivity of neurons in the anterior, posterior, and primary fields of cat auditory cortex. Hear. Res. *240*, 22–41.
16. Rajan, R., Aitkin, L.M., Irvine, D.R., and McKay, J. (1990). Azimuthal sensitivity of neurons in primary auditory cortex of cats. I. Types of sensitivity and the effects of variations in stimulus parameters. J. Neurophysiol. *64*, 872–887.
17. Stecker, G.C., Harrington, I.A., and Middlebrooks, J.C. (2005). Location coding by opponent neural populations in the auditory cortex. PLoS Biol. *3*, e78.
18. 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.
19. Middlebrooks, J.C., and Bremen, P. (2013). Spatial stream segregation by auditory cortical neurons. J. Neurosci. *33*, 10986–11001.
20. Maddox, R.K., Billimoria, C.P., Perrone, B.P., Shinn-Cunningham, B.G., and Sen, K. (2012). Competing sound sources reveal spatial effects in cortical processing. PLoS Biol. *10*, e1001319.
21. Edmonds, B.A., and Culling, J.F. (2005). The role of head-related time and level cues in the unmasking of speech in noise and competing speech. Acta Acust. U. Acust. *91*, 546–553.
22. Edmonds, B.A., and Culling, J.F. (2005). The spatial unmasking of speech: evidence for within-channel processing of interaural time delay. J. Acoust. Soc. Am. *117*, 3069–3078.
23. Thiran, A.B., and Clarke, S. (2003). Preserved use of spatial cues for sound segregation in a case of spatial deafness. Neuropsychologia *41*, 1254–1261.
24. Duffour-Nikolov, C., Tardif, E., Maeder, P., Thiran, A.B., Bloch, J., Frischknecht, R., and Clarke, S. (2012). Auditory spatial deficits following hemispheric lesions: dissociation of explicit and implicit processing. Neuropsychol. Rehabil. *22*, 674–696.
25. Lee, C.C., and Middlebrooks, J.C. (2011). Auditory cortex spatial sensitivity sharpens during task performance. Nat. Neurosci. *14*, 108–114.
26. van der Heijden, K., Rauschecker, J.P., Formisano, E., Valente, G., and de Gelder, B. (2018). Active sound localization sharpens spatial tuning in human primary auditory cortex. J. Neurosci. *38*, 8574–8587.
27. 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.
28. Fritz, J.B., Elhilali, M., and Shamma, S.A. (2007). Adaptive changes in cortical receptive fields induced by attention to complex sounds. J. Neurophysiol. *98*, 2337–2346.
29. 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.
30. van der Heijden, K., Rauschecker, J.P., de Gelder, B., and Formisano, E. (2019). Cortical mechanisms of spatial hearing. Nat. Rev. Neurosci. *20*, 609–623.
31. Zion Golumbic, E.M.Z., 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.
32. 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.
33. 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.
34. Wöstmann, M., Herrmann, B., Maess, B., and Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. Proc. Natl. Acad. Sci. USA *113*, 3873–3878.
35. 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.
36. Ray, S., and Maunsell, J.H.R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLOS Biol. *9*, e1000610.
37. Steinschneider, M., Liégeois-Chauvel, C., and Brugge, J.F. (2011). Auditory evoked potentials and their utility in the assessment of complex sound processing. In The Auditory Cortex (Springer), pp. 535–559.
38. Yang, X., Wang, K., and Shamma, S.A. (1992). Auditory representations of acoustic signals. IEEE Trans. Inf. Theor. *38*, 824–839.
39. Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature *403*, 309–312.
40. Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (1979). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010.
41. Chan, A.M., Dykstra, A.R., Jayaram, V., Leonard, M.K., Travis, K.E., Gygi, B., Baker, J.M., Eskandar, E., Hochberg, L.R., Halgren, E., et al. (2014). Speech-specific tuning of neurons in human superior temporal gyrus. Cereb. Cortex *24*, 2679–2693. https://doi.org/10.1093/cercor/bht127.
42. Ladefoged, P., and Johnson, K. (2014). A Course in Phonetics (Nelson Education).
43. Patel, J.K., Kapadia, C.H., and Owen, D.B. (1976). Handbook of Statistical Distributions (M. Dekker).
44. Rifkin, R., Yeo, G., and Poggio, T. (2003). Regularized least-squares classification. Nato Sci. Series Sub Series III Comput. Sys. Sci. *190*, 131–154.
45. Morosan, P., Rademacher, J., Palomero-Gallagher, N., and Zilles, K. (2005). Anatomical organization of the human auditory cortex:

cytoarchitecture and transmitter receptors. In The Auditory Cortex (Psychology Press), pp. 45–68.

46. Webster, D.B., and Fay, R.R. (2013). In The Mammalian Auditory Pathway: Neuroanatomy, *1* (Springer Science & Business Media).
47. Grothe, B., Pecka, M., and McAlpine, D. (2010). Mechanisms of sound localization in mammals. Physiol. Rev. *90*, 983–1012.
48. Miller, L.M., and Recanzone, G.H. (2009). Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. Proc. Natl. Acad. Sci. USA *106*, 5931–5935.
49. Lomber, S.G., and Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. Nat. Neurosci. *11*, 609–616.
50. 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.e5.
51. 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.
52. Buzsáki, G., Anastassiou, C.A., and Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. *13*, 407–420.
53. Lewald, J., Schlüter, M.C., and Getzmann, S. (2018). Cortical processing of location changes in a "cocktail-party" situation: spatial oddball effects on electrophysiological correlates of auditory selective attention. Hear. Res. *365*, 49–61.
54. Michalka, S.W., Kong, L., Rosen, M.L., Shinn-Cunningham, B.G., and Somers, D.C. (2015). Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks. Neuron *87*, 882–892.
55. Hill, K.T., and Miller, L.M. (2010). Auditory attentional control and selection during cocktail party listening. Cereb. Cortex *20*, 583–590.
56. O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., and Lalor, E.C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb. Cortex *25*, 1697–1706.
57. Ding, N., and Simon, J.Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. *107*, 78–89. https://doi.org/10.1152/jn.00297.2011.
58. Luo, Y., Chen, Z., and Mesgarani, N. (2018). Speaker-independent speech separation With deep attractor network. IEEE ACM Trans. Aud. Speech Lang. Process. *26*, 787–796. https://doi.org/10.1109/TASLP.2018.2795749.
59. Teoh, E.S., Ahmed, F., and Lalor, E.C. (2022). Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment. J. Neurosci. *42*, 682–691.
60. Kayser, C., Petkov, C.I., Lippert, M., and Logothetis, N.K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. Curr. Biol. *15*, 1943–1947.
61. 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.
62. Alho, K., Salmi, J., Koistinen, S., Salonen, O., and Rinne, T. (2015). Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks. Brain Res. *1626*, 136–145.
63. Shamma, S., and Fritz, J. (2014). Adaptive auditory computations. Curr. Opin. Neurobiol. *25*, 164–168.
64. Bagur, S., Averseng, M., Elgueda, D., David, S., Fritz, J., Yin, P., Shamma, S., Boubenec, Y., and Ostojic, S. (2018). Go/No-Go task engagement enhances population representation of target stimuli in primary auditory cortex. Nat. Commun. *9*, 2529.
65. Yin, P., Fritz, J.B., and Shamma, S.A. (2014). Rapid spectrotemporal plasticity in primary auditory cortex during behavior. J. Neurosci. *34*, 4396–4408.
66. 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.
67. 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.
68. Rauschecker, J.P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc. Natl. Acad. Sci. USA *97*, 11800–11806.
69. Rauschecker, J.P. (2018). Where, when, and how: are they all sensorimotor? Towards a unified view of the dorsal pathway in vision and audition. Cortex *98*, 262–268.
70. Alain, C., Arnott, S.R., Hevenor, S., Graham, S., and Grady, C.L. (2001). "What" and "where" in the human auditory system. Proc. Natl. Acad. Sci. USA *98*, 12301–12306.
71. Ahveninen, J., Jääskeläinen, I.P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., Levänen, S., Lin, F., Sams, M., Shinn-Cunningham, B.G., et al. (2006). Task-modulated "what" and "where" pathways in human auditory cortex. Proc. Natl. Acad. Sci. USA *103*, 14608–14613.
72. Romanski, L.M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P.S., and Rauschecker, J.P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat. Neurosci. *2*, 1131–1136.
73. Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J.P. (2001). Functional specialization in rhesus monkey auditory cortex. Science *292*, 290–293.
74. 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.
75. 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.
76. Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., and Leahy, R.M. (2011). Brainstorm: a user-friendly application for MEG/EEG analysis. Comput. Intell. Neurosci. *2011*, 879716.
77. Jot, J.-M., Larcher, V., and Warusfel, O. (1995). Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony. Journal of the Audio Engineering Society *98*, 3980.
78. Khalighinejad, B., Nagamine, T., Mehta, A., and Mesgarani, N. (2017). NAPLib: an open source toolbox for real-time and offline Neural Acoustic processing. In Proc. IEEE Int. Conf. Acoust. Speech Signal Process., *2017* (IEEE Publications), pp. 846–850.
79. Warusfel, O. (2003). In Listen HRTF Database (IRCAM and AK).
80. Crone, N.E., Boatman, D., Gordon, B., and Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. Clin. Neurophysiol. *112*, 565–582. https://doi.org/10.1016/S1388-2457(00)00545-9.
81. Edwards, E., Soltani, M., Kim, W., Dalal, S.S., Nagarajan, S.S., Berger, M.S., and Knight, R.T. (2009). Comparison of time–frequency responses and the event-related potential to auditory speech stimuli in human cortex. J. Neurophysiol. *102*, 377–386.
82. Chi, T., Ru, P., and Shamma, S.A. (2005). Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. *118*, 887–906. https://doi.org/10.1121/1.1945807.
83. Theunissen, F.E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network *12*, 289–316.
84. David, S.V., Mesgarani, N., and Shamma, S.A. (2007). Estimating sparse Spectro-temporal receptive fields with natural stimuli. Network *18*, 191–212.
85. 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. J. Neurosci. Methods *281*, 40–48.

86. 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.
87. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage *31*, 968–980.
88. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., and Collins, D.L.; Brain Development Cooperative Group (2011). Unbiased average age-appropriate atlases for pediatric studies. Neuroimage *54*, 313–327.
89. Yuan, J., and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. J. Acoust. Soc. Am. *123*, 3878.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MATLAB	MathWorks, Natick, MA	https://matlab.mathworks.com/
Freesurfer's parcellation	Dykstra et al.;[74] Fischl et al.[75]	https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki
Brainstorm	Tadel et al.[76]	https://neuroimage.usc.edu/brainstorm/Introduction
LISTEN HRTF	Jot et al., 1995[77]	http://recherche.ircam.fr/equipes/salles/listen/
Codes for phoneme analysis, high0gamma envelope extraction and STRF analysis	Khalighinejad et al.[78]	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805377/
Auditory spectrogram calculation	Neural Systems Laboratory	http://nsl.isr.umd.edu/downloads.html

RESOURCE AVAILABILITY

Lead contact

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).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All the original code and software used is publicly available as of the date of publication. DOIs are listed in the key resources table. The data that support the findings of this study are available upon request from the lead contact, Nima Mesgarani (nima@ee.columbia.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants and data collection

Seven adult subjects (four male and three female) with self-reported normal hearing and fluency in American English participated in this study. As a part of the clinical procedure to localize the epileptic focus, these participants were implanted with high-density multi-electrode arrays with contact spacing of 3.5mm (3 subjects bilaterally in the auditory cortex and 4 subjects unilaterally in the right auditory cortex). Intracranial EEG signals were recorded from these patients at a sampling rate of 3000 Hz per channel (16-bit precision, range ± 8 mV, DC) with a data acquisition module using RZ2 bioamp processor, PZ5 neurodigitizer amplifier and RS4 data streamer (Tucker-Davis Technologies (TDT), Alachua, FL, USA). Either subdural or skull electrodes were used as references, as dictated by recording quality at the bedside after online visualization of the signal. Speech signals were recorded in sync with the intracranial EEG for subsequent offline analysis. The research protocol was approved by the Institutional Review Board of Feinstein Institute for Medical Research, and signed consent forms were obtained from all participants in the study.

METHOD DETAILS

Stimulus

Narrative American English sentences (5s long on average) connected to form coherent stories were recorded in house from one male and one female talkers. These sentences were presented to subjects using standard Panasonic stereo in-ear headphones (Panasonic RP-HJE120) at a sampling rate of 44.1 KHz in the following three scenarios:

1 Single-talker task: The subjects listened to speech from a single male talker and a single female talker coming from the horizontal plane in two locations (45 and -45 degrees) in the azimuthal plane (Figure 1B). Fifty sentences divided equally between the two locations were spoken by each talker. The content of the story spoken by the male differed from that spoken by the female.

2 Multiple Spatially Separated Talkers task: The subjects listened to simultaneous speech from a male and a female talker separated in azimuthal space (45 and -45 degrees) (Figure 2A). The subjects listened to exactly the same mixture twice in two separate blocks. In the first block, they were asked to focus on the female talker while she randomly switched between the two azimuthal locations after each sentence; in the second block, they were asked to focus on the male talker as he randomly switched between the two azimuthal locations. For this case, we used the same 50 sentences from the single spatial speech task presented at the same locations, with the only modification being that the two talkers were simultaneously presented.
3 Multiple Spatially Overlapping Talkers: The same speech from the Multiple Spatially Separated Talkers scenario was played, but the talkers were collocated in all blocks. The blocks were structured in a similar manner to the Multiple Spatially Separated Talkers task, with two blocks for the same mixture speech, but subjects attended to the female in the first block and attended to the male in the second block.

To avoid confounding attentional effects, we presented the scenarios in the following order: 1) multiple spatially separated talkers 2) multiple spatially overlapping talkers and 3) single talker. Note that although sentences were repeated once in the multi-talker conditions (i.e. each sentence was presented once in the attended condition and once in the unattended condition), there was little chance of adaptation or familiarization. Specifically, attention was directed only once to each sentence and the large number of sentences (50 per speaker, 100 in total) introduced considerable time in between repetitions, reducing chances of adaptation and familiarization. We compared 30 individual HRTFs from the LISTEN HRTF dataset[79] and chose the HRTF filter of an average-size head in that dataset. We rendered the speech in azimuthal space by convolving the HRIR (Head Related Impulse Response) filters for the left and right ear with the mono speech sound to obtain a spatially rendered binaural speech at 0-degree elevation and +45-dgree and -45-degree azimuths.

QUANTIFICATION AND STATISTICAL ANALYSIS

Preprocessing neural data

Neural data were analyzed offline using MATLAB software (Mathworks Inc). All data were montaged to a common average reference[80] to reduce the effect of recording noise. The data were resampled to 400 Hz. Line noise at 60 Hz and its harmonics (up to 180 Hz) were removed using a notch filter (designed using function *fir2* and implemented using function *filtfilt*) with an order of 1000. The envelope of the high gamma band (70-150 Hz) was then extracted from the data using the Hilbert transform followed by downsampling the data to 100 Hz.[81] To obtain the envelope of this broad-band signal, we first filtered the data into eight frequency bands (10Hz wide each) between 70 and 150 Hz using MATLAB's fft and ifft functions- code available as EcogExtractHighGamma.m.[78] After this, the envelope of each band was obtained by taking the absolute value of the Hilbert transform and finally the envelope of all the eight frequency bands were summed to compute the final envelope. We normalized the neural responses to zero mean and unit standard deviation. The normalization was done by finding the mean and standard deviation from male and female blocks combined (but separately done for single and multi-talker experiments). We selected responsive electrodes by measuring the STRF prediction correlation for all electrodes in the single-talker condition. We defined responsive electrodes as the electrodes that had significant STRF prediction correlations, resulting in 119 electrodes.

Calculating the spectrotemporal receptive fields (STRF)

For all the STRF analysis, we used mono-stimulus (stimulus with no spatial information) to estimate the encoding of spectrotemporal features. We first converted the sound waveform onto a time-frequency representation using a model of cochlear frequency analysis, consisting of a bank of 128 asymmetric constant-Q filters equally spaced on a logarithmic axis.[82] The Matlab code to calculate the auditory spectrogram is available at http://nsl.isr.umd.edu/downloads.html. The output of the filter bank was then resampled to 50 frequency bands to prevent parameter overfitting. We calculated the STRF of each electrode using the normalized reverse correlation algorithm (STRFLab software package available at http://www.strflab.berkeley.edu).[83] Regularization and 4-fold cross-validation techniques were used to prevent overfitting of the STRF.[84] Sparseness values are picked from 8, 16, 32 and tolerance values from 0.5, 0.05, 0.1. Note that the algorithm picks the exact same parameters (sparseness=8, tolerance=0.05) for the analysis in Figure 5 making the STRFs for attend male vs female comparable. The response latency parameter was estimated by finding the centroid of the excitatory region of the STRF along the time dimension (single-talker STRFs were used for this estimation).

To obtain the STRF prediction correlations (analysis in Figures 1D and 2B), we first predicted the neural response from an electrode's nonspatial STRF using a 4-fold cross-validation. We then correlated the predicted neural response to the actual neural response to obtain a STRF prediction correlation value. The STRF measured for analysis in Figure 5 used the mixture speech stimulus. For each electrode, we first measured STRF separately in each of the two attention conditions from Figure 5A (attend female and attend male) using the exact same mixture speech. We then averaged the STRFs across electrodes to obtain a mean attend-female STRF and mean-attend male STRF. We subtracted the resultant attend-male STRF from attend-female STRF to calculate the modulatory effects of attention to talker on spectrotemporal receptive fields. These effects hold regardless of the location of attended talker as shown in Figure S3. STRF analysis from Figure 5 has been repeated and verified using stereo STRFs in Figure S3.

Brain maps

The electrodes were mapped onto the brain of each subject using coregistration by iELVis[85] followed by their identification on the postimplantation CT scan using BioImage Suite.[86] Following coregistration, electrodes were snapped to the closest point on the reconstructed brain surface of the preimplantation MRI. We used the FreeSurfer automated cortical parcellation[74,75] to identify the anatomical regions in which each electrode contact was located within ~3 mm resolution. We used Destrieux's parcellation,[35] which provides higher specificity in the ventral and lateral aspects of the medial temporal lobe compared to the Desikan Killiany atlas.[87] These automated parcellation labels were closely inspected by the neurosurgeon using the subject's coregistered postimplant MRI. The electrodes were plotted on the average brain template ICBM152[88] using Brainstorm.[76]

Estimation of the metrics of CSI, CSI*attend*, location selectivity, AMI-location, talker selectivity, AMI-talker, and ST

Calculation of single-talker contralateral selectivity index (CSI)

Using the neural responses during the single-talker task, we measured the location selectivity of an electrode from the mean-response level (MRL) of an electrode to define a contralateral selectivity index (CSI):

$$MRL_{45} = \frac{1}{T}\sum_{t=0}^{T} r_{45}(t)\ MRL_{-45} = \frac{1}{T}\sum_{t=0}^{T} r_{-45}(t)$$

$$CSI = \frac{MRL_{-45} - MRL_{45}}{MRL_{-45} + MRL_{45}}$$

Where $r_{45}(t)$ and $r_{-45}(t)$ are the unnormalized high gamma power envelope (microvolts squared) response of each electrode to the speech arriving from 45 and -45 degrees respectively. For each location, we concatenated all trials over time excluding the initial 0.5s of each trial to limit the analysis to the sustained parts of the response as done in.[12] T is the total duration of the stimulus in each location (after excluding the first 0.5s from each trial). Location selectivity for each electrode is measured as the absolute value of CSI.

*Calculation of CSI*attend *and AMI-Location*

For multi-talker, we quantified the modulation of the neural responses with attention to location as follows:

$$MRL\_attend_{45} = \frac{1}{T}\sum_{t=0}^{T} r\_attend_{45}(t)\ MRL\_attend_{-45} = \frac{1}{T}\sum_{t=0}^{T} r\_attend_{-45}(t)$$

$$CSI_{attend} = \frac{MRL\_attend_{-45} - MRL\_attend_{45}}{MRL\_attend_{-45} + MRL\_attend_{45}}$$

Where $r\_attend_{45}(t)$ and $r\_attend_{-45}(t)$ denote the unnormalized neural responses of each electrode to the mixed speech when the subject attends to talker at 45 and -45 degrees (calculated similarly as described in section above), respectively, and CSI_{attend} quantifies the modulation of neural responses with attention to 45 and -45 degree locations. Finally, the modulation of the contralateral selectivity index with attention to location, $AM_{direction}$, is defined as the absolute value of CSI*attend*:

$$AM_{direction} = |CSI_{attend}|$$

Calculation of single-talker Talker Selectivity

We measured talker selectivity for an electrode by performing a two-sample t-test between the sustained neural response to all single-talker male speech and the sustained neural responses to all single-talker female speech (500ms after the trial onset to the end of trial). We then computed the absolute of the t-value of this t-test to measure the preference of the electrode responds to one talker over the other.[3]

Calculation of AMI-Talker

Our definition of attention modulation to the talker, similar to the definition in,[3] was as follows:

$$AM_{Talker} = \frac{\overline{Crr}_{attend} - \overline{Crr}_{unattend}}{1 + \overline{Crr}_{attend} + \overline{Crr}_{unattend}}$$

where $\overline{Crr}$ in this equation denotes the average correlation coefficient between the actual and STRF predicted neural responses (average is taken across trials). The predicted neural responses used in the calculation of $\overline{Crr}_{attend}$ and $\overline{Crr}_{unattend}$ are obtained from the STRFs that are calculated from the same neural response to mixed speech but with respect to single-talker spectrograms of attended and unattended talkers, accordingly. As a result, the $\overline{Crr}_{attend}$ and $\overline{Crr}_{unattend}$ for a given mixture measure the encoding of attended and unattended speech in the neural responses. Therefore, the electrodes that are modulated more by attention to a talker will have higher values of $\overline{Crr}_{attend}$ than the $\overline{Crr}_{unattend}$ and consequently higher values of AM_{Talker}.

To determine the significance of $AM_{direction}$, we generated a null distribution for AM_{Talker} using the same equation above, but the STRF prediction values are instead calculated by correlating the shuffled predicted neural response along the trials to the actual

neural response. Similar to[3] for AMI-talker analysis (Figure 5), we used only the electrodes whose AMI was higher than the threshold established for statistical significance from the null distribution (i.e., AMI > 0.04, p<0.05). This resulted in 44/119 (36.97%) electrodes from our study.

Calculation of Spatial Tuning (ST)

We defined ST for each electrode as follows:

$$ST = \frac{\overline{Crr}_{contralateral} - \overline{Crr}_{ipsilateral}}{1 + \overline{Crr}_{contralateral} + \overline{Crr}_{ipsilateral}}$$

Where $\overline{Crr}$ in this equation represents the average correlation coefficient between actual and STRF predicted neural responses (average is taken across trials). The predicted neural responses used in the calculation of $\overline{Crr}_{contralateral}$ and $\overline{Crr}_{ipsilateral}$ are obtained from the STRFs that are calculated from the same neural response to mixed speech but with respect to the single-talker spectrograms of talkers in contralateral and ipsilateral locations relative to the hemispheric location of each electrode. For a given spatially separated mixture stimulus, $\overline{Crr}_{contralateral}$and $\overline{Crr}_{ipsilateral}$ measure the encoding of contralateral and ipsilateral speech in neural response. Therefore, the electrodes that become more selective to contralateral or ipsilateral locations have a higher difference between these values and therefore, higher absolute ST values

Phoneme F-Statistics and decoding analysis

To quantify phonemic information that is represented in the neural responses, we measured the separability of responses to 46 individual phonemes by calculating the f-statistics between the responses to these phonemes. We aligned the neural responses in each task to the corresponding phoneme onsets in the single-talker stimulus and segmented the neural response using the phoneme transcription. Phoneme transcription was obtained by segmenting the single-talker stimulus into time-aligned sequences of phonemes using the Penn Phonetics Lab Forced Aligner Toolkit.[89] We grouped the segmented phoneme responses into consonants and vowels resulting in a total of 335 consonants and 335 vowels in each of the 4 conditions shown in Figures 1B and 2A for the qualitative plots. For the quantitative plots, we then calculated the f-statistics of the neural responses to 46 individual phonemes as a ratio of between-class to within-class variability.[43]

For the electrode population decoding, we used a regularized least square (RLS)[44] linear classifier to decode 46 individual phonemes from the neural data by training on 90% data and testing on 10% over 10 cross-validations. We trained the classifier separately for each time point after the onset of the phoneme to obtain phoneme decoding accuracies over time.

Calculating the acoustic difference between talkers

We measured the time frequency representations of the single male and single female talker speech using the same cochlear model described above.[82] We then collapse the time dimension of this representation by averaging the spectrograms over time separately for the male and female talkers to get the average spectral profile for each talker. We then subtracted the male and the female spectral profile for the analysis in Figure 5.

Linear regression analysis

We used linear regression to fit a model to determine how well the single-talker tuning parameters predict the multi-talker tuning parameters (we used *fitlm* function). The p-values and the confidence intervals for the model weights were calculated using cross-validation and were used to estimate the contribution of each single-talker parameter.