Joint population coding and temporal coherence link an attended talker's voice and location features in naturalistic multi-talker scenes

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

Affiliations:

1. Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, Netherlands.
2. Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, United States.
3. Maastricht Center for Systems Biology (MaCSBio), Maastricht University, Maastricht, Netherlands.
4. Department of Electrical Engineering, Columbia University, New York, United States.
5. Hofstra Northwell School of Medicine, New York City, United States.
6. The Feinstein Institute for Medical Research, New York City, United States.

HIGHLIGHTS

- Cortical responses to an single talker exhibit a distributed gradient, ranging from sites that are sensitive to both a talker's voice and location (dual-feature sensitive sites) to sites that are sensitive to either voice or location (single-feature sensitive sites).
- Population response patterns of dual-feature sensitive sites encode voice and location features of the attended talker in multi-talker scenes jointly and with equal precision.
- Despite their sensitivity to a single feature at the level of individual cortical sites, population response patterns of single-feature sensitive sites also encode location and voice features of a talker jointly, but with higher precision for the feature they are primarily sensitive to.
- Neural sites which selectively track an attended speech stream concurrently encode the attended talker's voice and location features.
- Attention selectively enhances temporal coherence between voice and location selective sites over time.
- Joint population coding as well as temporal coherence mechanisms underlie distributed multi-dimensional auditory object encoding in auditory cortex.

ABSTRACT (240 words)

Listeners readily extract multi-dimensional auditory objects such as a 'localized talker' from complex acoustic scenes with multiple talkers. Yet, the neural mechanisms underlying simultaneous encoding and

linking of different sound features – for example, a talker's voice and location – are poorly understood. We analyzed invasive intracranial recordings in neurosurgical patients attending to a localized talker in real-life cocktail party scenarios. We found that sensitivity to an individual talker's voice and location features was distributed throughout auditory cortex and that neural sites exhibited a gradient from sensitivity to a single feature to joint sensitivity to both features. On a population level, cortical response patterns of both dual-feature sensitive sites but also single-feature sensitive sites revealed simultaneous encoding of an attended talker's voice and location features. However, for single-feature sensitive sites, the representation of the primary feature was more precise. Further, sites which selective tracked an attended speech stream concurrently encoded an attended talker's voice and location features, indicating that such sites combine selective tracking of an attended auditory object with encoding of the object's features. Finally, we found that attending a localized talker selectively enhanced temporal coherence between single-feature voice sensitive sites and single-feature location sensitive sites, providing an additional mechanism for linking voice and location in multi-talker scenes. These results demonstrate that a talker's voice and location features are linked during multi-dimensional object formation in naturalistic multi-talker scenes by joint population coding as well as by temporal coherence between neural sites.

SIGNIFICANCE STATEMENT

Listeners effortlessly extract auditory objects from complex acoustic scenes consisting of multiple sound sources in naturalistic, spatial sound scenes. Yet, how the brain links different sound features to form a multi-dimensional auditory object is poorly understood. We investigated how neural responses encode and integrate an attended talker's voice and location features in spatial multi-talker sound scenes to elucidate which neural mechanisms underlie simultaneous encoding and linking of different auditory features. Our results show that joint population coding as well as temporal coherence mechanisms contribute to distributed multi-dimensional auditory object encoding. These findings shed new light on cortical functional specialization and multidimensional auditory object formation in complex, naturalistic listening scenes.

INTRODUCTION

In everyday life, listeners rapidly and effortlessly parse complex acoustic scenes with multiple sound sources into its individual constituents. This process of auditory scene analysis (ASA[1]) is based on the segregation and subsequent grouping of features of temporally overlapping sound sources, resulting in the formation of coherent auditory objects[2]. Sound features contributing to auditory object formation include voice features related to object identity (e.g., pitch or timbre) and location features (e.g.,

interaural time differences, location cues)[3–5]. However, the neural basis for multi-dimensional auditory object formation in complex, naturalistic listening scenes is poorly understood.

One unresolved question is how cortical representations of individual sound features are linked by the brain to form a multi-dimensional auditory object. If voice and location features are encoded independently in two separate, functionally specialized and hierarchical processing streams as posited by the prevailing dual-stream framework[6,7], it is not clear how these features are subsequently integrated to form a multi-dimensional auditory object. In contrast, recent studies using an active task design indicate that sound feature encoding may be distributed across auditory cortex rather than taking place in dedicated, functionally specialized anatomical regions as posited by the dual-stream theory. For example, studies in cats[8] and humans[9] showed that spatial sensitivity in primary auditory cortex (PAC) sharpens during goal-directed sound localization, suggesting that regions that are not considered part of the location pathway (i.e. PAC) may be recruited flexibly for spatial processing based on behavioral goals. Additionally, while speech processing has been attributed mostly to posterior STG[10,11], a recent study demonstrated that speech processing is instead distributed across auditory cortex[12]. Such findings indicate that sound (feature) encoding may be more distributed than posited by the hierarchical dual-stream framework.

Additionally, it is not understood what neural mechanisms integrate cortical representations of individual sound features (e.g. spatial and non-spatial features). One hypothesis is that neuronal populations are sensitive to specific combinations of features and thereby encode multiple dimensions of an auditory object. Prior studies confirmed that some cortical sites are sensitive to multiple sound features simultaneously (e.g. in ferrets[13], for a review[14]). However, because most prior measurements were performed with single sound sources, it is not known whether these cortical sites maintain their multi-dimensional sensitivity when presented with complex acoustic scenes comprising multiple, interfering sound sources. An alternative hypothesis states that auditory streams (pertaining to auditory objects) are formed through temporal coherence, i.e., response synchronization between neural populations that are sensitive to specific sound features[15]. Neural measurements in animals[16,17] and humans[18] demonstrate that temporal coherence is a plausible mechanism for auditory feature binding and segregation. It remains to be evaluated whether temporal coherence also underlies linking of voice and location features in human auditory cortex in naturalistic listening scenes.

Finally, although it is well known that auditory attention modulates the neural representation of spatial and non-spatial features[19,20] as well as auditory object formation[21,22], it is not known how attention modulates integrated encoding of spatial and non-spatial features in complex, naturalistic sound scenes. Moreover, it remains an open debate[2] whether auditory objects form pre-attentively[23] or whether attention is necessary for auditory object formation[15].

Here, we investigated cortical multi-dimensional auditory object formation with stereotactic electroencephalography (sEEG) recordings in neurosurgical patients. We measured neural activity in response to real-world sound scenes consisting of a single localized talker or two spatially separated talkers. The unique spatiotemporal resolution of neurophysiological recordings enabled us to map feature encoding and multi-dimensional object formation across auditory cortex. We found that active listening to complex, naturalistic scenes gives rise to distributed but joint voice and location encoding in single- as well as in multi-talker scenes. Furthermore, our results revealed that response patterns of distinct neural populations jointly encoded an attended talker's voice and location features. Finally, we show that attending to a localized talker in multi-talker scenes selectively enhanced temporal coherence between voice and location sensitive sites. In sum, these data demonstrate that multiple neural mechanisms contribute to linking an attended talker's voice and location in multi-talker scenes. .

RESULTS

We analyzed neural measurements in seven neurosurgical patients recorded with intracranial depth electrodes (stereoelectroencephalography, sEEG; Methods). Participants listened to English speech utterances consisting of one or two spatialized talkers. In single-talker scenes, either a male or female talker was present at a location of -45° or +45°. In two-talker scenes, a male and female talker were simultaneously present, one at -45° and the other at +45° (Figure 1 A). Trials had an average duration of 5 s and the location of the talkers changed at random after each trial. The total duration of each condition (i.e., single-talker speech and multi-talker speech) was 8 minutes. For the single-talker condition, speech was paused at random intervals between trials and the participant was asked to repeat the last sentence as well as the location of the talker. For the multi-talker condition, participants were instructed at the start of a block to attend to a specific talker (i.e. 'attend male' or 'attend female'). At random moments in between trials, participants were asked to report the location of the attended talker and the last sentence uttered by the attended talker. Participants successfully performed the behavioral task (see [24] for a detailed analysis of the behavioral results).

Cortical sensitivity to a talker's voice and location features

We observed significant neural population responses to speech in the high gamma envelope of 147 cortical sites in auditory cortex (paired samples t-test of responses to speech versus silence, $p < 0.05$, FDR corrected, $q < 0.05$; Figure 1 D). These speech responsive sites were located in Heschl's gyrus (HG, 6 left hemisphere, 32 right hemisphere), planum temporale (PT, 11 left hemisphere, 24 right hemisphere) and superior temporal gyrus (STG, 25 left hemisphere, 49 right hemisphere).

We characterized response properties for voice and location features by examining the responses to the single talker scenes for each cortical site. To assess to what extent a site exhibited sensitivity to voice, to

location, or to both, we contrasted the responses to one class of a feature (e.g., the male voice) to the responses to the other class of the feature (e.g., the female voice). For all sites, we extracted the mean response for each trial as the mean from 0.5 s post sound onset to 1.5 s post sound onset (that is, excluding the onset response). Figure 1 B shows example neural responses of three sites: One site sensitive to voice features (top panels), one site sensitive to location features (middle panels) and one site sensitive to both voice and location features (bottom panels). Figure 1 C shows the resulting response distributions for the sites in Figure 1 B. To test for sensitivity to voice features, we computed the effect size (Cohen's d[21]) for the difference between the mean responses to all male and female trials, irrespective of the location of the talker (50 trials each). To test for sensitivity to location features, we computed Cohen's d for the difference between the mean responses to all trials in which the talker was at the right and all trials in which the talker was at the left, irrespective of the talker's voice (50 trials each). Figure 1 D depicts voice and location sensitivity (Cohen's d) on the cortical surface. There was no overall relationship between sensitivity strength for a single talker's voice and location features (|Cohen's d|, $r = 0.037$, $p = 0.66$).

Statistical testing confirmed that 47 sites were significantly sensitive to voice features only (paired samples t-tests, $p < 0.05$, FDR corrected) and 12 sites were significantly sensitive to location features only (paired samples t-tests, $p < 0.05$, FDR corrected). In agreement with prior results (e.g. [24,25]), most sites which were sensitive to location, preferred locations in the contralateral hemifield. Further, 23 sites were sensitive to both location and voice features ($p < 0.05$ for both t-tests). While multi-dimensional sensitivity has only been demonstrated for combinations of non-spatial features in humans, these results confirm prior work in animals[26] which showed that some neuronal populations in auditory cortex are sensitive for both spatial and non-spatial features[13]. In sum, cortical responses reveal a gradient from single-feature voice or location sensitive sites to dual-feature voice and location sensitive sites.

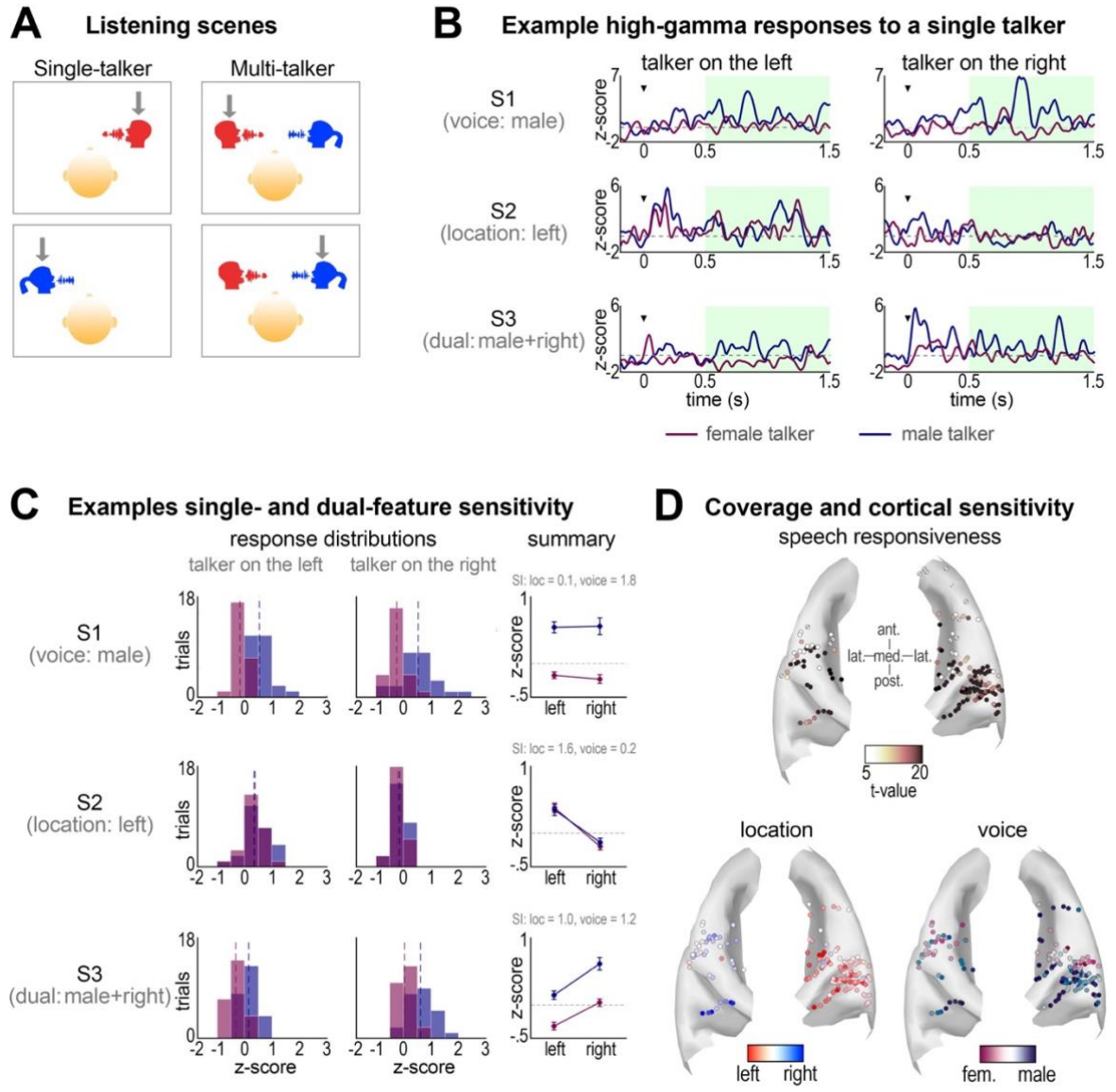

Figure 1. Experiment design and single-talker cortical sensitivity for location and voice features. (A) Two examples of single-talker scenes (left panels) and two examples of spatial multi-talker scenes (right panels). Gray arrows indicate the attended talker. (B) Example neural responses from three sites: a single-feature voice sensitive site (S1, higher responses to male talker than to female talker irrespective of location), a single-feature location sensitive site (S2, higher responses to a talker on the left than a talker on the right, irrespective of the talker), and a dual-feature sensitive site (S3, higher responses to a male talker than to a female talker and higher responses to a talker on the right than a talker on the left). Black triangle indicates sound onset. Shaded green area depicts the time window for calculating voice and location sensitivity (i.e., 500 – 1,500 ms post sound onset). (C) Distribution of average trial responses to the male and female talker for three example sites (same as in B). Dashed line indicates the median of each distribution. Panels on the right depict the mean and standard error of the mean for each distribution. The sensitivity index (SI) is the effect size of the difference in response to two locations (SI loc) or the difference in response to two talkers (SI voice). (D) Top panel: Speech responsiveness of all electrodes in AC. Color saturation reflects the *t*-value for the contrast speech versus silence (see Methods). Electrodes that did not exhibit a significant response to speech are indicated by a slanted black line. Lower panels: Sensitivity for a single talker's location (left panel)

and voice features (right panel) plotted on the cortical surface for all speech responsive sites. Color indicates Cohen's *d* (range [-1,1]).

Spectrotemporal tuning properties explain sensitivity to a talker's voice and location features

Prior work showed that spectrotemporal tuning properties explain preferential responses to a talker's voice[21,27]. We examined whether we observe a similar relationship between spectrotemporal tuning and sensitivity to a talker's voice features in the present dataset and, additionally, we examine to what extent spectrotemporal tuning properties can also explain sensitivity to a talker's location features. We characterize the spectrotemporal tuning properties of each speech responsive site by estimating a spectrotemporal receptive field (STRF) from the responses to single-talker stimuli. We estimated STRFs using a five-fold cross-validation procedure, leaving out 20 trials and fitting the STRF on the remaining 80 trials. We used the left-out 20 trials to estimate the goodness of fit, calculating the correlation between these left-out neural responses and neural responses predicted by the fitted STRFs (Methods). Next, we examined to what extent STRF shape explained sensitivity to talker's voice and location features for all cortical sites with a well-fitted STRF (correlation $r > 0.2$, $n = 93$).

To analyze the relationship between STRF shape and sensitivity to a talker's voice features, we divided the group of sites sensitive to a talker's voice ($n = 47$, Figure 1) into sites responding maximally to the male talker and sites responding maximally to the female talker. In line with prior work[21], the average STRF of sites responding preferentially to the female talker exhibited tuning properties corresponding to the spectral profile of the female talker, while the average STRF across sites responding preferentially to the 'male' talker exhibited tuning properties corresponding to the spectral profile of the male talker. That is, Figure 2 A shows that the average STRF of 'male'-preferring sites exhibited an excitatory region at low frequencies between 50 Hz and 100 HZ, overlapping with F0 of the male talker (65 Hz). In contrast, the average STRF of 'female'-preferring sites exhibited an excitatory region between 160 Hz and 200 Hz, overlapping with F0 of the female talker (175 Hz).

To quantify this relationship between spectral tuning properties and sensitivity to a talker's voice features, we extracted the spectral receptive fields (SRFs) of sites responding maximally to the male talker and sites responding maximally to the female talker. The SRF corresponds to the first component of a principal component analysis (PCA) of the STRF along the spectral dimension[21]. The difference in the SRFs of these two groups (i.e., responding preferentially to the female or male voice) was strongly correlated to the difference between the spectral profile of the male and female talker, indicating that the preference of voice sensitive sites for the male or female talker was driven by the correspondence between the spectral response profile of the site and the acoustic profile of the talker ($r = 0.85$, $p = 7.1E-15$; Methods; Fig. 2 B). Further, mapping the SRFs to sensitivity for a talker's voice using ridge regression

(Methods), showed that SRFs predicted sensitivity for a talker's voice well. That is, there was a high correlation between predicted sensitivity and observed sensitivity (r = 0.932, p = 5.9E=17; Fig. 2 C). These findings confirm that sensitivity to a talker's voice is driven by the spectral tuning properties of cortical sites[21].

We then repeated the STRF analysis to assess the relationship between STRF shape and sensitivity to a talker's location features. First, we computed the average STRF across sites that are sensitive to a talker's location (n = 12, Figure 1) and across sites that were not sensitive to a talker's location (and not sensitive to a talker's voice either, n = 65). Figure 2 D shows that sites which were sensitive to a talker's location had an excitatory STRF region for frequencies above 1.5 kHz which was reduced in sites which were not sensitive to a talker's location. In contrast, sites which were sensitive to a talker's location responded more weakly to frequencies between 0.5 – 1.5 kHz. No difference in STRF properties was observed for frequencies below 0.5 kHz.

As expected, extracting and comparing the SRFs of the two groups (i.e., location-sensitive versus not location-sensitive) showed that the difference in SRFs was not correlated to the difference between the spectral profile of the male and female talker (r = 0.17, p = 0.23; Methods; Fig. 2 D). However, mapping SRFs to sensitivity to a talker's location features (Methods), demonstrated that SRFs predicted such location-sensitivity well: There was a high correlation between predicted and observed sensitivity (r = 0.932, p = 5.9E=17; Fig. 2 C). These findings indicate that sites which are sensitive to a talker's location respond more strongly to frequencies with robust interaural level difference (ILD) cues for sound localization[28], while responding less strongly to frequencies in which binaural disparity cues such as ILDs and interaural time differences (ITDs) are less reliable[29]. Further, these results indicate that sensitivity to a talker's location is not related to tuning to low frequencies (i.e., < 0.5 kHz). Taken together, these findings confirm previous work demonstrating that spectrotemporal tuning explains tuning to a talker's voice[21,27] and extend this by showing that spectrotemporal tuning also explains tuning to a talker's location.

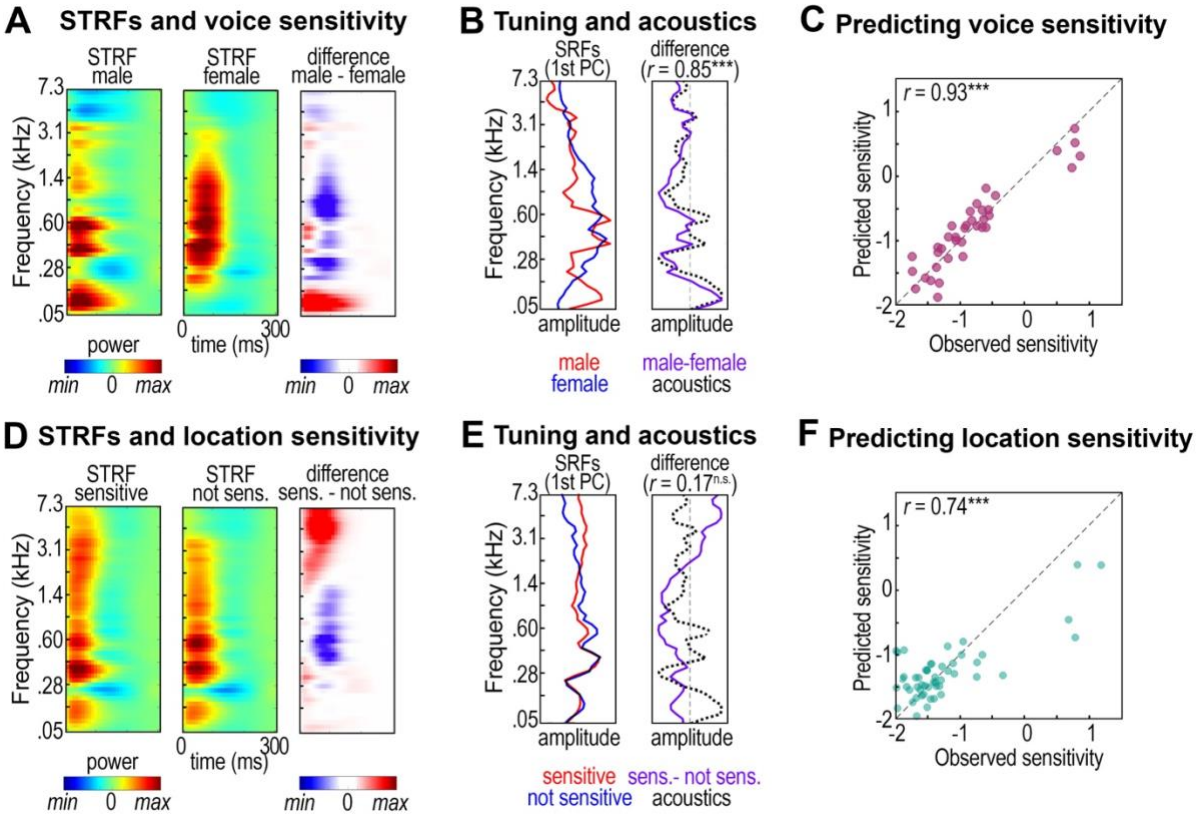

Figure 2. Spectrotemporal tuning characteristics explain sensitivity to a talker's voice and to a talker's location. (A) Spectrotemporal tuning properties related to voice sensitivity. Form left to right: Average STRF for sites responding maximally to the male talker, average STRF for sites responding maximally to the female talker and the difference (STRF male – STRF female). (B) Comparing spectral tuning properties to the acoustics of the male and female talker. Left panel: Average spectral receptive field of sites responding maximally to a female talker (blue). Right panel: The correlation between the difference SRF (SRF male – SRF female) and the difference in the acoustics of the male and female talker. (C) Predicting voice sensitivity from the difference SRF (Cohen's *d*). (D) Spectrotemporal tuning properties related to location sensitivity. From left to right: Average STRF for location sensitive sites, average STRF for sites that were not sensitive to location and the difference (STRF sensitive – not sensitive). (E) Comparing spectral tuning properties to location sensitivity. Left panel: Average spectral receptive field of sites sensitive to location features (red) and for sites not sensitive to location features (blue). Right panel: No correlation between the difference SRF (SRF not location sensitive – SRF location sensitive) and the difference in the acoustics of the male and female talker. (F) Predicting location sensitivity (|Cohen's *d*|) from the difference SRF.

Sensitivity to a talker's voice and location across the cortical hierarchy

To investigate to what extent sensitivity to a talker's voice and location can be related to cortical processing stages, we investigated how sensitivity to a talker's features was distributed across auditory cortex. While several studies linked delineated anatomical regions to hierarchical processing stages (for example, HG is considered primary auditory cortex and PT and STG higher-order auditory regions[30]), other work investigating neural response latencies and response properties showed that a single anatomical region may contain different auditory processing stages (e.g. [12,31]). That is, as response

latency roughly corresponds to the number of synapses away from the periphery it is considered as an indication of the processing stage of a neural site. Here, we therefore assessed the distribution of feature sensitivity both within cortical auditory regions and as a function of response latency. We calculated response latency as the peak along the temporal dimension of the STRF (for sites with a well-fitted STRF, $r > 0.2$, $n = 93$; Methods).

Figure 3 A shows the regional distributions of Cohen's *d* for a talker's voice. Comparing the distributions showed that sensitivity to a talker's voice was stronger in HG than in STG (|Cohen's d|, Kruskal-Wallis H test, $\chi^2(2) = 14.6$, $p = 0.0007$, Figure 3 A). Further, there was a negative correlation between sensitivity to a talker's voice and response latency ($r = -0.526$, $p = 1.3E-7$; Figure 3 B). These findings confirm prior reports of a decrease in sensitivity to a talker's voice along the cortical auditory processing hierarchy[21]. In contrast, although we observed a trend towards regional differences in the distribution of Cohen's *d* for a talker's location (|Cohen's *d*|, Kruskal-Wallis H test, $\chi^2(2) = 5.45$, $p = 0.07$; Figure 3 A), this trend failed to reach significance. Moreover, we observed no correlation between sensitivity to a talker's location and response latency ($r = -0.146$, $p = 0.16$; Figure 3 B). While the lack of regional differences may be a consequence of the relatively low anatomical sampling density, together the regional and response latency results indicate that sensitivity is consistent across low- and high-level processing stages during active listening. These findings confirm recent work[9,24], but contrast the predictions of the dual-stream framework which posits that PT is functionally specialized for spatial processing[6,7,32].

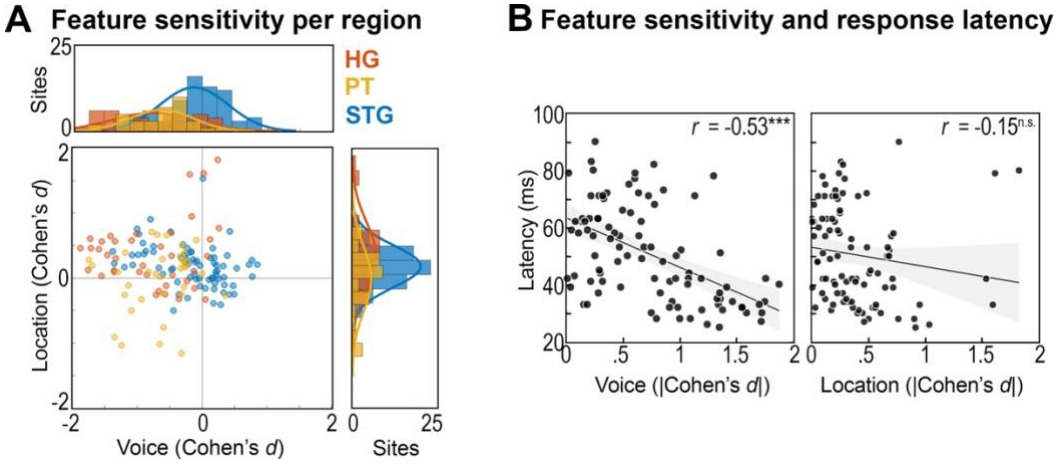

Figure 3. Sensitivity to a single talker's voice and location across the cortical hierarchy (A) Scatterplot of voice sensitivity (x-axis) and location sensitivity (y-axis). Each symbol represents an individual site. Bar graphs depict corresponding marginal distributions for voice sensitivity (left) and location sensitivity (right). (B) Correlation between single-talker response latency and feature sensitivity (left panel: voice; right panel: location). Each circle depicts a site. Solid lines depict the correlation; shaded areas depict the 95% confidence interval. Asterisks indicate significance: *** = $p < 0.001$.

Attentional modulation of neural responses to a talker's voice and location in multi-talker scenes

We showed that cortical sites exhibit varying degrees of sensitivity for a single talker's location and voice. Motivated by prior findings of attentional modulation of neural responses to a talker's voice and location[21,24], we examined to what degree we observed such local attentional modulations in multi-talker scenes in our data. Further, we characterized how attentional modulation by a talker's voice relates to attention modulation by a talker's location. Specifically, we quantified to what extent attending to a talker's voice and location in multi-talker scenes modulated the response gain of individual cortical sites similar to our quantification of sensitivity to a single talker's voice and location features. Specifically, we calculated the effect size Cohen's *d* for the difference in mean response to the trials for each attentional condition. As before, we calculated the mean response for each trial from 0.5 s post sound onset to 1.5 s post sound onset (excluding the onset response).

In agreement with prior studies[21,24], attending a localized talker evoked weak response gain modulations across speech responsive sites both by the attended talker's voice and by the attended talker's location. Figure 4 A shows that attentional modulation of response gain was smaller than modulation by a single talker's voice or location, that is, single-talker sensitivity (paired samples t-test of |Cohen's *d*|; voice: $t(146) = 9.65$, $p = 2.38E\text{-}17$; location: $t(146) = 6.68$, $p = 4.64E\text{-}10$). In agreement with this, statistical testing did not identify neural sites of which the response gain was modulated significantly by attention to the talker's voice (paired samples t-tests, $p > 0.05$), the talker's location (paired samples t-tests, $p > 0.05$), or jointly ($p > 0.05$). Further, Figure 4 B shows that only few sites were jointly modulated by an attended talker's voice and location in multi-talker scenes. Specifically, only few electrodes were close to the diagonal and exhibited attentional response gain modulation for both voice and location (|Cohen's *d*| > .1). Crucially, as single-source sensitivity to a specific sound feature (e.g. pitch, location) is generally considered an indication of functionally specialized processing[2,32], the lack of corresponding, dedicated attentional modulations of response gains raises the question what the role is of these sites in the encoding of an attended sound source in scenes with multiple sound sources.

Decoding a localized talker from population activity patterns in multi-talker scenes

To elucidate the relationship between local encoding properties and population encoding properties, we examined whether a localized talker can be decoded from population response patterns. Specifically, we used a linear decoding approach to assess to what extent a localized talker can be decoded from population responses in single talker scenes and to what extent an attended localized talker can be decoded from population responses in multi-talker scenes. To decode a localized talker in single-talker scenes, we trained a four-class regularized least-squares (RLS[22,33]) classifier on the response patterns in single-talker scenes using a leave-two-trials-out cross-validation procedure (corresponding to 25 folds).

To decode an attended localized talker, we trained an identical four-class RLS classifier on the response patterns in multi-talker scenes using a similar cross-validation procedure. We assessed decoding accuracy by predicting the talker's voice and location from the response patterns of the left-out trials of each fold (Methods).

As expected, Figure 4 C shows that a single localized talker could be accurately decoded from the entire population of speech responsive sites (n = 147; average accuracy [standard error of the mean; SEM] = 93.0 % [2.29], p = 0, FDR corrected). Similarly, the attended localized talker was decoded accurately from the entire population of speech responsive sites (mean accuracy [SEM] = 63.0 % [5.80], p = 0). Marginal decoding accuracies for the talker's voice and location show that both features were decoded with equal precision in single-talker scenes (Figure 4 D, mean marginal accuracy: voice [SEM] = 97.0 % [1.66], p = 0; location [SEM] = 96.0 % [1.87], p = 0; paired samples t-test, $t(24)$ = 0.37, p = 0.72) as well as in multi-talker scenes (Figure 3D, mean accuracy voice [SEM] = 82.0 % [4.68], p = 0; average accuracy location [SEM] = 74.0 % [3.95], p = 0; $t(24)$ = 1.69, p = 0.14). These findings show that although attentional modulation of local response gain by the attended talker's voice and location in multi-talker scenes was weak (Figure 4 A), response patterns across the entire population of speech responsive sites the attended localized talker with high fidelity.

Next, we examined how sites which exhibit single-feature sensitivity for a talker's voice or location features in their local responses (n = 47 and n = 12, Figure 1) encode a localized talker in population response patterns. We therefore trained the RLS classifier on the population responses of these neural sites with the same procedure described above. Note that although the latter population is relatively small, we chose to use this stringent selection to ensure that the population did not incorporate sites that were also to some extent sensitive to a talker's voice. Figure 4 C shows that the classifier successfully decoded a localized talker in single-talker scenes from voice sensitive sites (mean accuracy [SEM] = 65.0 % [4.33], p = 0) as well as from location sensitive sites (mean accuracy 61.0 % [4.10], p = 0). Further, the marginal accuracies in Figure 4 D show that the classifier decoded the talker's voice more accurately from population responses of voice sensitive electrodes than the talker's location (mean marginal accuracy: voice [SEM] = 99.0 % [1.00], p = 0; location [SEM] = 65.0 % [4.33], p = 0.0098; paired samples t-test, $t(24)$ = 7.49, p = 3.93E-7). Conversely, decoding accuracy was higher for the talker's location than for the talker's voice when the classifier operated on population responses of location sensitive sites (voice [SEM] = 65.0 % [4.33], p = 0.02; location [SEM] = 91.0 % [2.45], p = 0.0024; $t(24)$ = 5.32, p = 3.74E-5). These findings show that single-feature sensitive sites nevertheless encode coarse information about other feature dimensions of an individual talker in their population responses.

Furthermore, in multi-talker scenes, the classifier decoded the attended localized talker above chance level from population responses of voice sensitive sites (Figure 4 C, mean accuracy [SEM] = 55.0 % [5.59],

p = 0) as well as from population responses of location sensitive sites (Figure 4 C, mean accuracy [SEM] = 36.0 % [3.84], p = 0.026). However, while both the attended talker's location and voice were decoded above chance level from population responses of voice sensitive sites (mean marginal accuracy: voice [SEM] = 80.0 % [5.40], p = 0 ; location [SEM] = 63.0 % [4.59], p = 0.023), only the attended talker's location was decoded accurately from population responses of location sensitive sites. Specifically, the decoding accuracy for the attended talker's voice just failed to reach statistical significance, which may be a consequence of the small number of sites in this group (mean marginal accuracy: voice [SEM] = 54.0 % [2.77], p = 0.079; location [SEM] = 67.0 % [3.45], p = 0.021). For both voice and location sensitive sites, the preferred feature was decoded significantly better than the other feature (voice: $t(24)$ = 2.72, p = 0.024; location: $t(24)$ = 2.98, p = 0.024). These findings indicate that populations which exhibit local properties of functional specialization in response to single-source sound scenes may nonetheless encode (coarse) information about multiple dimensions of the auditory object. Future work including more fine-grained sampling of multiple feature dimensions (e.g., more talkers and more voices) is required to establish the resolution with which population responses of single-feature sensitive sites encode other feature dimensions.

Finally, we showed previously that some neural sites were sensitive both for a talker's voice and location (Figure 1, n = 23). We examined to what extent population responses of these dual-feature sensitive sites also jointly encode a talker's voice and location in single- and in multi-talker conditions. That is, while prior work in animals showed that auditory cortex contains sites which are sensitive to spatial as well as non-spatial sound features[13], work in humans focused only on multi-dimensional sensitivity for non-spatial features (e.g. [26]). Moreover, as all prior studies were conducted with single-source scenes, it is not clear to what extent multi-dimensional sensitivity is maintained in multi-talker sites. Here, Figure 4 C shows that the localized talker was decoded accurately from population responses of dual-feature sensitive sites in single-talker scenes (mean accuracy [SEM] = 91.0 % [3.50], p = 0). The attended localized talker was also decoded accurately from population responses in multi-talker scenes (mean accuracy [SEM] = 47.0 % [3.63], p = 0). Importantly, Figure 4 D shows that dual-feature sensitive encoded the talker's voice and location with equal precision, both in single-talker (mean marginal accuracy: voice [SEM] = 96.0 % [1.87], p = 0; location [SEM] = 94.0 % [2.61], p = 0; paired samples t-test: $t(24)$ = 0.81, p = 0.57) and in multi-talker scenes (mean marginal accuracy: voice [SEM] = 63.0 % [3.57], p = 0.006; location [SEM] = 69.0 % [3.62], p = 0; paired samples t-test: $t(24)$ = 1.1, p = 0.28). In sum, we show that population responses of dual-feature sensitive sites encode both spatial and non-spatial features of an attended talker in multi-talker scenes. This suggests that such dual-feature sensitive sites contribute to the encoding of multiple dimensions of an auditory object.

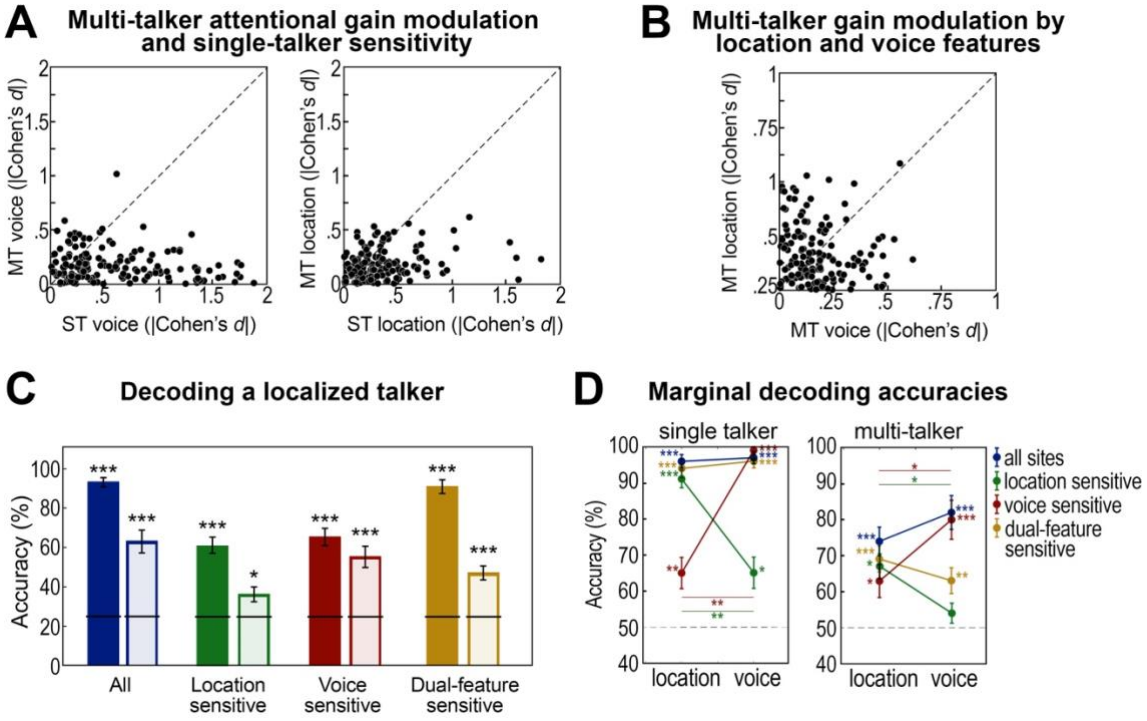

Figure 4. Attentional gain modulation and the representation of a localized talker in single- and multi-talker scenes. (A) Relationship between attentional gain modulation and single-talker sensitivity. Left panel: Gain modulation by attending to a talker's voice in multi-talker scenes (y-axis) versus single-talker voice sensitivity (x-axis). Right panel: Gain modulation by attending a talker's location in multi-talker scenes (y-axis) versus single-talker location sensitivity (x-axis). (B) Scatterplot of gain modulation by an attended talker's voice (x-axis, |Cohen's *d*|) and by an attended talker's location (y-axis, |Cohen's *d*|). (C) Decoding a localized talker from response patterns in single-talker scenes (filled bars) and decoding an attended localized talker in multi-talker scenes (open bars). Horizontal lines depict chance level. Asterisks indicate significance: *** = $p < 0.001$. (D) Marginal decoding accuracies for a talker's voice and location. Dashed line depicts chance level, asterisks indicate significance: * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$.

Sites selectively tracking an attended speech stream simultaneously encode an attended talker's voice and location features

In the preceding sections, we examined to what extent single-feature sensitive and dual-feature sensitive sites encode an attended talker's voice and location features in multi-talker scenes. However, prior work showed that auditory cortex also contains neural sites which are not strongly sensitive to a single talker's features, but which nonetheless play a crucial role in auditory object formation by selectively tracking the attended speech stream in multi-talker listening scenes[21,22,24,34]. As the relationship between such speech stream tracking and encoding of the attended talker's features is not known, we analyzed the measured neural responses to multi-talker scenes to evaluate to what degree such neural sites which selectively track sites additionally encode the attended talker's voice and location features.

For each site, we first quantified selective tracking of the attended speech stream by calculating to what extent a site's responses in spatial multi-talker scenes were modulated by attention to reflect the response to the attended localized talker in single talker scenes. We define the tracking index (TI) for each site similar to the definition in [21]:

$$Tracking\ Index = corr(att_M, single_M) - corr(att_M, single_F) + corr(att_F, single_F) - corr(att_F, single_M)$$

Here, M refers to the male talker and F to the female talker. Further, $corr(att, single)$ corresponds to the correlation between the single-talker response and the multi-talker response for the same trial calculated over the entire duration over the trial (see example in Figure 5 A).

Next, to gain more insight into the encoding properties of sites which selectively track an attended speech stream, we examined to what extent TI is explained by attentional modulation of STRF properties[35,36]. We quantified such attentional modulation of STRF properties by estimating for each cortical site two STRFs from the responses to multi-talker scenes: one for the 'attend male talker' condition and one for the 'attend female talker' condition. To relate attention-induced spectrotemporal plasticity to the encoding of a talker's spectral characteristics, we extracted the spectral receptive field (SRF) from each STRF as the first principal component of a PCA (only for cortical sites with a robust STRF as estimated from single-talker responses, $r > 0.2$, $n = 93$) and compared these to the spectral profile of the male and female talker. Specifically, we computed the difference in SRF for the two attention conditions (i.e., attend female – attend male) and correlated this difference SRF to the acoustic difference spectrum between the male and female talker (see examples in Figure 5 B, C). If a site's spectral tuning properties are modulated by attention towards the attended talker's spectral profile, we expect a high correlation between the difference SRF and the acoustic difference spectrum (Fig. 5 C). That is, we expect these sites to resemble a contrast matched filter which facilitates figure-ground segregation by enhancing the attended target (e.g., the female talker) and filtering out the background (e.g., the male talker)[37]. Therefore, we quantified the strength of attentional modulation of STRF properties by calculating a contrast matched filter (CMF) index, which is the correlation between the attention-driven difference in the SRFs and the acoustic difference spectrum for the female and male talker (Figure 5 C).

As expected[24], CMF explains TI well ($r = 0.577$, $p = 1.4E\text{-}9$, Fig. 5 D). This indicates that sites whose STRF properties are strongly modulated by attention tend to be sites which selectively track the attended speech stream. In contrast, TI is not correlated to single talker encoding properties (voice sensitivity: $r = 0.11$, $p = 0.24$; location sensitivity: $r = 0.12$, $p = 0.24$) or multi-talker attentional response gain modulation (attended talker's voice: $r = 0.06$, $p = 0.48$; attended talker's location: $r = -0.13$, $p = 0.24$). Thus, the encoding properties of sites which selectively track an attended speech stream are

characterized by attentional modulation of STRF properties rather than by single-talker sensitivity or multi-talker attentional response gain modulation.

Further, to examine whether the population responses of sites which selectively track the attended speech stream also encode the attended talker's voice and location, we trained the four-class classifier on their population response patterns in multi-talker scenes (i.e, for sites with TI > 0.1, n = 33). The classifier accurately decoded the attended localized talker from these population response patterns (average accuracy [SEM] = 51.0 % [5.10], p = 0; Figure 5 E). Furthermore, the classifier decoded the attended talker's voice and location with equal precision (Figure 5 E; marginal accuracies: voice [SEM] = 72.0 % [3.63], p = 0; location [SEM] = 70.0 % [4.56], p = 0; paired samples t-test, $t(24)$ = 0.40, p = 0.69). In sum, population responses of sites which selectively tracked the attended speech stream also encoded the attended talker's voice and location. This finding indicates that the population responses of these sites play a role in combining selective tracking of an attended auditory object (here, speech stream) with encoding of the features of that object (here, the talker's voice and location).

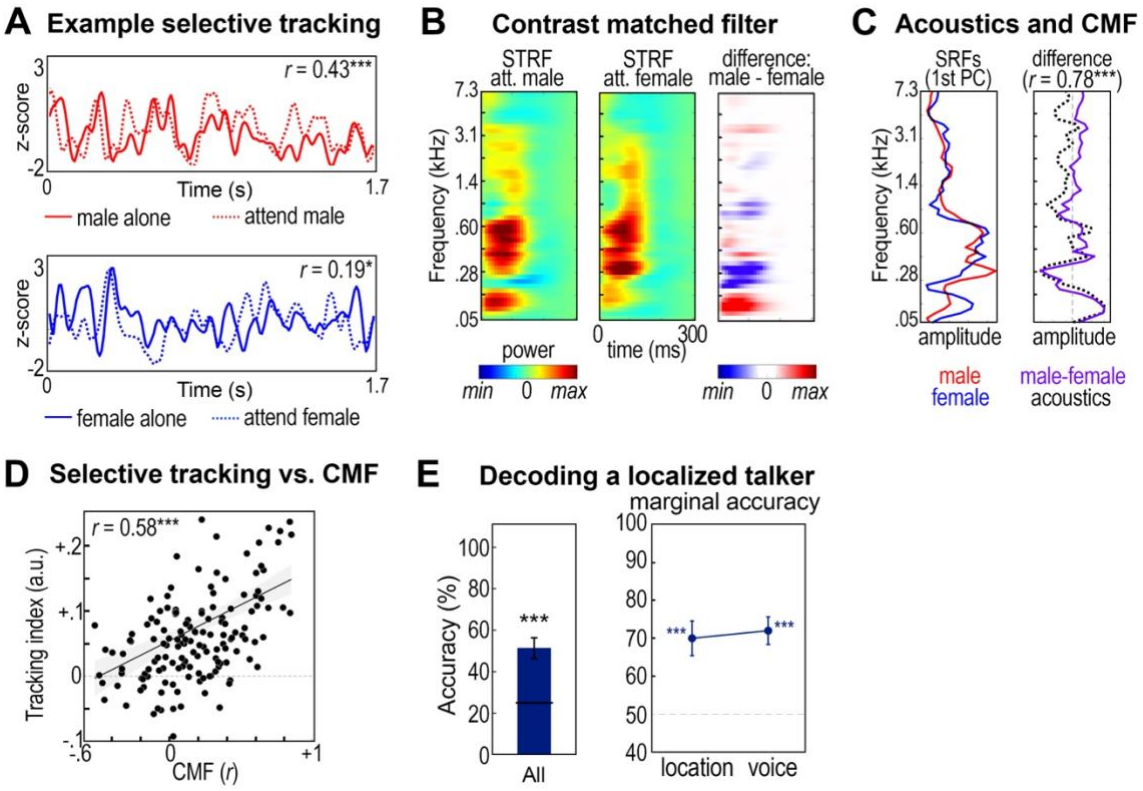

Figure 5. Selective speech tracking and encoding of the attended talker's voice and location. (A) High-gamma responses of an example site exhibiting selective tracking of the attended talker. Solid lines depict response in single-talker scene, dotted lines depict response in multi-talker scene. (B) STRFs of an example cortical site exhibiting contrast matched filtering. Left panel: STRF in the 'attend male' condition. Middle panel: STRF in the 'attend female' condition. Right panel: Difference (STRF attend male – STRF attend female). (C) Comparing spectral tuning properties in the two attention conditions to the acoustics of the male and female talker. Left panel: Spectral receptive fields

for the 'attend male' condition (red) and for the 'attend female' condition (blue). Right panel: The correlation between the difference SRF (SRF attend male – SRF attend female) and the difference in the acoustics of the male and female (D) Correlation between CMF (x-axis) and Tracking Index (y-axis). Circles represent cortical sites. (E) Left panel: Decoding an attended localized talker from population responses in multi-talker scenes. Horizontal line depicts chance level. Asterisks indicate significance: *** = $p < 0.001$. Right panel: Marginal decoding accuracies for a talker's voice and location. Dashed line depicts chance level, asterisks indicate significance: * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$.

Attention to a localized talker enhances temporal coherence between voice-sensitive and location-sensitive sites.

We showed that joint population coding results in simultaneous encoding of an attended talker's voice and location in spatial multi-talker scenes. However, other mechanisms may also contribute to linking an attended talker's voice and location features in spatial multi-talker scenes. In particular, it has been proposed that different sound features are bound together through synchronization of the slow fluctuations in neural responses of feature sensitive cortical sites, that is, temporal coherence[15]. Here, we evaluated to what extent such temporal coherence contributed to linking the attended talker's voice and location. First, we computed temporal coherence between the high-gamma envelope of pairs of neural sites consisting of one site sensitive to a single talker's voice and one site sensitive only to a single talker's location (Figure 6 A). For each voice-location site pair ($n = 57$), we quantified temporal coherence of the high gamma envelope at frequencies between 2-22 Hz using the coherency coefficient. The coherency coefficient is the frequency-domain mathematical equivalent of the cross-correlation function in the time-domain[38] (within-subjects analysis; Methods). As shown in Figure 6 A, we evaluated the hypothesis that attention selectively enhances temporal coherence between the voice and location site in each pair by contrasting temporal coherence in different attention conditions. That is, we examined whether temporal coherence increased when attention was directed towards a localized talker which matched the pair's preferred features (condition 'preferred features attended') in comparison to when attention was directed to a localized talker which was orthogonal to the pair's preferred features (condition 'preferred features unattended'). For example, for a pair of sites consisting of a voice sensitive site tuned to the *female talker* and a location sensitive site tuned to the *right*, we hypothesize that temporal coherence increases when attention is directed to a female talker on the right in comparison to when attention is directed to a male talker on the left and the preferred features are therefore unattended (Figure 6 A). We quantify such attentional modulation of temporal coherence as the coherence gain:

$$AM_{coh} = \frac{coh_{xy}(\omega)_{attended} - coh_{xy}(\omega)_{unattended}}{coh_{xy}(\omega)_{unattended}}$$

Here, we define 'attended' as the condition in which attention is directed towards the preferred features and the 'unattended condition' as the condition in which the preferred features are unattended (i.e., because attention is directed towards orthogonal features).

First, we evaluated the development of attentional modulation of temporal coherence over time by calculating broadband temporal coherence (i.e. across all frequencies between 2 and 22 Hz) in shifting windows of 1,000 ms with 50 % overlap. Figure 6 B shows a weak but not statistically significant attentional enhancement of temporal coherence immediately post sound onset (t-test, p = 0.074, FDR corrected) and a strong and robust gain in temporal coherence starting at approximately 2 s post stimulus-onset. Furthermore, we examined whether the observed attentional enhancement in the late response was generic to the range of frequencies tested here (2-22Hz) or whether it was frequency specific. We therefore repeated the analysis on the late response which showed the most robust attentional temporal coherence gain (i.e., from 1.75 s until 3.25 s post sound onset) in narrow frequency bins of 3 Hz (central frequencies [CF]: 3, 6, 9, 12, 15, 18, 21 Hz). We found that in this time window, attentional enhancement of temporal coherence was generic for frequencies < 22 Hz (Figure 6 C).

Taken together, these results demonstrate that in spatial multi-talker scenes, attention selectively enhanced temporal coherence between sites sensitive to a single talker's voice and sites sensitive to a single talker's location. Moreover, we showed that this attentional enhancement builds up over time. In sum, temporal coherence is a plausible binding mechanism for linking voice and location encoding by single-feature sensitive sites in order to form a complete auditory object in complex, multi-source auditory scenes.

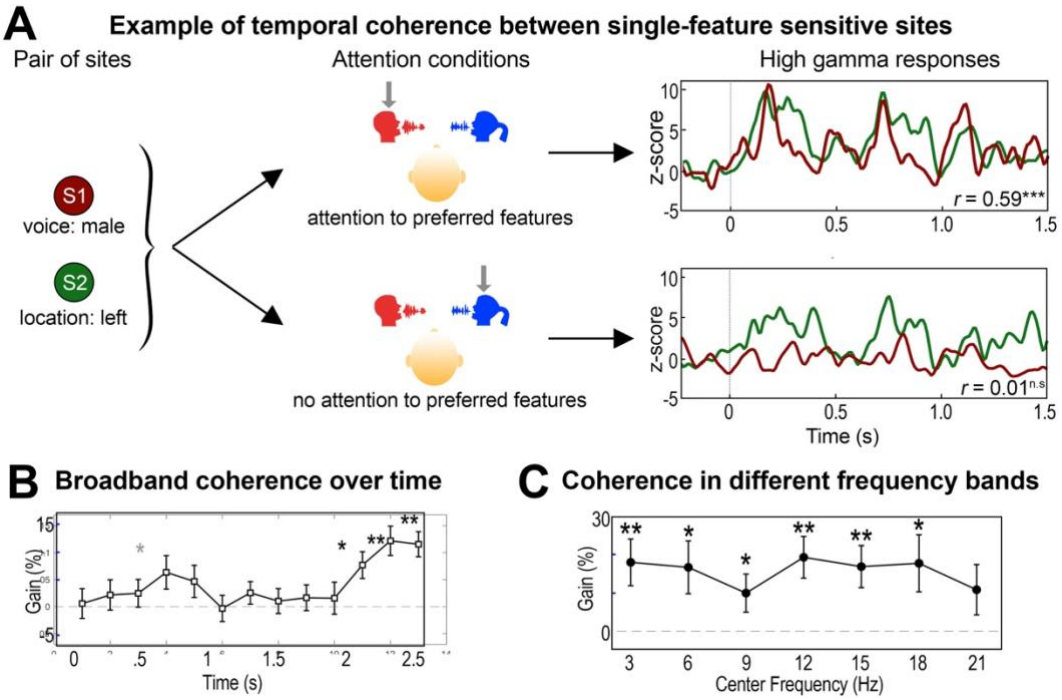

Figure 6: Linking an attended talker's voice and location through temporal coherence. (A) Schematic example of attentional modulation of temporal coherence of a pair of neural sites consisting of a single-feature voice sensitive site (red) and a single-feature location sensitive site (green). (B) Development of broadband temporal coherence gain over time. Error bars reflect standard error of the mean (SEM). Asterisks indicate a significant attentional enhancement of coherence. ** = $p < 0.01$, * = $p < 0.05$. (C) Temporal coherence gain per narrowband frequency bin.

DISCUSSION

In daily-life situations, listeners flexibly extract relevant information from cluttered and dynamic auditory scenes to form multi-dimensional auditory objects such as a 'localized talker'. While auditory object formation is critically dependent on the integration of different feature dimensions (e.g. location and voice), it is presently not clear how such different sound attributes are linked by the brain. Here, we utilized the unique spatiotemporal resolution of invasive intracranial measurements in neurosurgical patients to gain insight into the neural mechanisms linking voice and location sound features in real-life listening scenes consisting of a single talker or two spatially separated talkers.

We found that cortical responses varied from dual-feature sensitivity to a talker's voice and location, to single-feature sensitivity to a talker's voice or location only. Further, population responses of both dual-feature sensitive and single-feature sensitive sites, simultaneously encoded an attended talker's voice and location features. Our findings thus indicate that cortical representations of a multi-dimensional localized talker are derived from joint encoding in distributed population response patterns rather than separate voice and location encoding in dual processing streams within delineated anatomical regions[6,7,32]. Furthermore, our data indicate that attention enhances temporal coherence between voice and location sensitive sites, thereby providing an additional mechanism for linking the representations of both features. These results provide important new insights into the emergence of multi-dimensional auditory objects[2,4] in auditory cortex during active, goal-oriented listening in real-life listening scenes.

Active task design and naturalistic stimuli reveal distributed voice and location encoding

Our data showed that the sensitivity of local cortical sites for a talker's voice and location features can be explained by the underlying spectrotemporal tuning properties. These results align with prior research attributing speaker sensitivity to spectrotemporal tuning properties[21] and fast temporal processing to the posterior-dorsal regions of human auditory cortex[25] which tend to show strong spatial sensitivity[25]. Additionally, our results highlight that voice and location encoding during active listening occurs in distributed networks that span the entire auditory cortex rather than within clearly delineated, functionally specialized cortical regions. Moreover, linking local responses to population encoding showed that sites which are characterized by functionally specialized local responses (for example, voice

sensitive sites), nevertheless encode information about both voice and location in their population responses.

Further, the distributed networks of voice and location sensitivity that we observed at the level of individual cortical sites is in agreement with a recent study which demonstrated that acoustic and phonetic processing in auditory cortex are based on distributed, parallel processing rather than serial processing[12]. This indicates that distributed processing may be a general characteristic of auditory encoding and speech encoding specifically[10]. Moreover, the occurrence of distributed voice and location representations as observed in the present study conceivably ensures sufficient flexibility to accommodate sound encoding in changing acoustic environments and with changing behavioral goals[39].

Further, we showed that sensitivity to a talker's location features is similar across sites that are at lower stages of the hierarchy and sites that are at higher stages of the hierarchy. These results deviate from the view that spatial sensitivity emerges only in higher-order regions belonging to the functionally specialized location stream[7,40]. Instead, our findings are in agreement with more recent studies with active task designs which demonstrated that neural location sensitivity in early processing stages (i.e. primary auditory cortex) is more pronounced during active, goal-oriented localization[8,9]. Taken together, our results emphasize that experiment designs comprising active tasks and naturalistic stimuli are crucial to uncover representational mechanisms related to goal-oriented behavior in complex auditory scenes.

Pre-attentive and attentive linking of voice and location to form complete auditory objects

Whether attention is required for auditory object formation remains a matter of debate[2,15]. Some have argued that auditory streams are formed pre-attentively, for example by the activation of separate populations of neurons[23]. Others have posited that attention is required to bind together the various attributes of the attended object[15]. Our data showed that a subset of local cortical sites exhibited sensitivity to both voice and location features, similar to prior findings of multi-feature sensitivity in ferret auditory cortex[13]. Moreover, we showed that the population response patterns of these sites gave rise to representations of the multi-dimensional object, that is, the localized talker. It is therefore conceivable that the activation of these populations contributes to stream formation in the spatial multi-talker scenes utilized here. However, to what extent this mechanism is pre-attentive requires further investigation.

Our results also revealed top-down attentional modulation of feature binding. That is, a subset of sites showed single-feature sensitive responses to either voice or location features, in agreement with the 'feature analysis'-stage of the temporal coherence framework. According to this framework, distinct neural populations generate representations of various sound properties[15,41,42]. Here, we found that attention selectively enhanced temporal coherence between relevant single-feature voice and location

sensitive sites. This result is consistent with accumulating evidence[16,18] supporting temporal coherence as a potential mechanism for grouping of perceptual features. An open question is where in the cortex the read-out of such temporally coherent input takes place.

Taken together, our data indicate that linking of a talker's voice and location features in spatial multi-talker scenes emerges from a mixture of (potentially pre-attentive) activation of dual-feature sensitive neural sites, population coding and attentional modulation of temporal coherence between voice and location sensitive sites.

A continuum of attentional modulations of voice and location encoding

In agreement with prior work[21,24], our results show that attending to a talker's voice and location elicited weak attentional response gain control. Further, attention dynamically changed spectrotemporal tuning properties of late-response cortical sites, resulting in contrast matched filtering shape changes that enhanced local selective tracking of the attended talker's speech. These results connect prior work in animals which showed that task performance and attention changed spectrotemporal tuning in auditory cortex[36,37] to attended speech encoding in complex scenes in human auditory cortex. Moreover, these results extend findings from prior neural measurements in human auditory cortex, which showed that contextual information elicited adaptive STRF tuning to boost perception of degraded speech[35].

Further, an attended talker's location features elicited comparable attentional gain control in early- and late-response sites, suggesting that attention affects spatial processing at low-level as well as higher-order processing stages. More research is needed to establish whether these spatial attention effects in low-level cortical regions emerge from feedback projections originating in higher-order regions[39]. Additional work is also needed to evaluate whether attending a talker's location in multi-talker scenes affects spatial tuning. That is, studies using single-source experiment designs with an active listening task reported sharpening of spatial tuning in primary auditory regions[8,9] and it is likely that similar effects take place in multi-talker scenes to support segregating background from foreground. However, to assess this hypothesis, an experiment design with more fine-grained sampling of azimuth locations is required to elucidate attentional modulation of spatial receptive fields.

Conclusion and outlook

Our results point to distributed and joint voice and location encoding across auditory cortex during active, goal-directed behavior. These findings support the view that object formation and attentional selection emerge gradually and in a distributed manner from the auditory hierarchy, rather than at one specific site or region in auditory cortex[4]. Such a distributed code flexibly accommodates rapid changes in the (acoustic) environment as well as changing behavioral goals. Crucially, the present findings demonstrate the need for real-life, complex stimuli and experimental designs including active behavioral

tasks to understand cortical processing of multi-dimensional auditory objects. Future studies including stimuli spanning a larger and more fine-grained range of talkers, locations and other sound features can further unravel local cortical tuning properties as well as population representations of multi-dimensional auditory objects. Finally, complementing sEEG measurements with high-density intracranial measurements (e.g. high-density electrocorticography [ECoG], e.g. [43]) are critical to refine cortical maps of local feature sensitivity, to tease apart fine-grained population representations within and across macro-anatomical regions, and to further our insights into feature binding through temporal coherence.

REFERENCES

1. Bregman, A. S. *Auditory Scene Analysis: The Perceptual Organization of Sound*. (MIT press, 1994).
2. Bizley, J. K. & Cohen, Y. E. The what, where and how of auditory-object perception. *Nat Rev Neurosci* **14**, 693–707 (2013).
3. Darwin, C. J. Auditory grouping. *Trends Cogn Sci* **1**, 327–333 (1997).
4. Shinn-Cunningham, B., Best, V. & Lee, A. K. C. Auditory object formation and selection. in *The auditory system at the cocktail party* 7–40 (Springer, 2017).
5. Bizley, J. K., Walker, K. M. M., Nodal, F. R., King, A. J. & Schnupp, J. W. H. Auditory cortex represents both pitch judgments and the corresponding acoustic cues. *Current Biology* **23**, 620–625 (2013).
6. Tian, B., Reser, D., Durham, A., Kustov, A. & Rauschecker, J. P. Functional specialization in rhesus monkey auditory cortex. *Science (1979)* **292**, 290–293 (2001).
7. Rauschecker, J. P. & Tian, B. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences* **97**, 11800–11806 (2000).
8. Lee, C.-C. & Middlebrooks, J. C. Auditory cortex spatial sensitivity sharpens during task performance. *Nat Neurosci* **14**, 108–114 (2011).
9. van der Heijden, K., Rauschecker, J. P., Formisano, E., Valente, G. & de Gelder, B. Active sound localization sharpens spatial tuning in human primary auditory cortex. *Journal of Neuroscience* **38**, 8574–8587 (2018).
10. Bhaya-Grossman, I. & Chang, E. F. Speech computations of the human superior temporal gyrus. *Annu Rev Psychol* **73**, 79–102 (2022).
11. Yi, H. G., Leonard, M. K. & Chang, E. F. The encoding of speech sounds in the superior temporal gyrus. *Neuron* **102**, 1096–1110 (2019).
12. Hamilton, L. S., Oganian, Y., Hall, J. & Chang, E. F. Parallel and distributed encoding of speech across human auditory cortex. *Cell* **184**, 4626–4639 (2021).
13. Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J. & Schnupp, J. W. H. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *Journal of neuroscience* **29**, 2064–2075 (2009).
14. King, A. J., Teki, S. & Willmore, B. D. B. Recent advances in understanding the auditory cortex. *F1000Res* **7**, (2018).
15. Shamma, S. A., Elhilali, M. & Micheyl, C. Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* **34**, 114–123 (2011).
16. Lu, K. *et al.* Temporal coherence structure rapidly shapes neuronal interactions. *Nat Commun* **8**, 1–12 (2017).
17. Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J. & Shamma, S. A. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* **61**, 317–329 (2009).

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

19. King, A. J. & Walker, K. M. M. Listening in complex acoustic scenes. *Curr Opin Physiol* **18**, 63–72 (2020).

20. Fritz, J. B., Elhilali, M., David, S. V & Shamma, S. A. Auditory attention—focusing the searchlight on sound. *Curr Opin Neurobiol* **17**, 437–455 (2007).

21. O'Sullivan, J. *et al.* Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron* **104**, 1195–1209 (2019).

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

23. Sussman, E. S., Horváth, J., Winkler, I. & Orr, M. The role of attention in the formation of auditory streams. *Percept Psychophys* **69**, 136–152 (2007).

24. Patel, P. *et al.* Interaction of bottom-up and top-down neural mechanisms in spatial multi-talker speech perception. *Current Biology* **32**, 3971–3986 (2022).

25. Derey, K., Valente, G., de Gelder, B. & Formisano, E. Opponent coding of sound location (azimuth) in planum temporale is robust to sound-level variations. *Cerebral Cortex* **26**, 450–464 (2016).

26. Allen, E. J., Burton, P. C., Olman, C. A. & Oxenham, A. J. Representations of pitch and timbre variation in human auditory cortex. *Journal of neuroscience* **37**, 1284–1293 (2017).

27. Steinschneider, M., Nourski, K. V, Rhone, A. E., Kawasaki, H. & Oya, H. Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. *Front Neurosci* **8**, 103289 (2014).

28. Schnupp, J., Nelken, I. & King, A. *Auditory Neuroscience: Making Sense of Sound*. (MIT press, 2011).

29. Risoud, M. *et al.* Sound source localization. *Eur Ann Otorhinolaryngol Head Neck Dis* **135**, 259–264 (2018).

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

31. Nourski, K. V *et al.* Functional organization of human auditory cortex: investigation of response latencies through direct recordings. *Neuroimage* **101**, 598–609 (2014).

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

33. Rifkin, R., Yeo, G. & Poggio, T. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences* **190**, 131–154 (2003).

34. Golumbic, E. M. Z. *et al.* Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **77**, 980–991 (2013).

35. Holdgraf, C. R. *et al.* Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nat Commun* **7**, 13654 (2016).

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

37. Fritz, J. B., Elhilali, M., David, S. V & Shamma, S. A. Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? *Hear Res* **229**, 186–203 (2007).

38. Bastos, A. M. & Schoffelen, J.-M. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front Syst Neurosci* **9**, 175 (2016).

39. van der Heijden, K., Rauschecker, J. P., de Gelder, B. & Formisano, E. Cortical mechanisms of spatial hearing. *Nat Rev Neurosci* **20**, 609–623 (2019).

40. Alain, C., Arnott, S. R., Hevenor, S., Graham, S. & Grady, C. L. "What" and "where" in the human auditory system. *Proceedings of the national academy of sciences* **98**, 12301–12306 (2001).

41. Fishman, Y. I., Arezzo, J. C. & Steinschneider, M. Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *J Acoust Soc Am* **116**, 1656–1670 (2004).

42. Fishman, Y. I., Reser, D. H., Arezzo, J. C. & Steinschneider, M. Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hear Res* **151**, 167–187 (2001).

43. Ramsey, N. F. *et al.* Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. *Neuroimage* **180**, 301–311 (2018).

44. Kayser, C., Petkov, C. I. & Logothetis, N. K. Tuning to sound frequency in auditory field potentials. *J Neurophysiol* **98**, 1806–1809 (2007).

45. Nir, Y. *et al.* Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Current biology* **17**, 1275–1285 (2007).

46. Yang, X., Wang, K. & Shamma, S. A. Auditory representations of acoustic signals. *IEEE Trans Inf Theory* **38**, 824–839 (1992).

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

48. Theunissen, F. E. *et al.* Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems* **12**, 289 (2001).

49. Santoro, R. *et al.* Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput Biol* **10**, e1003412 (2014).

50. Cooke, M. A glimpsing model of speech perception in noise. *J Acoust Soc Am* **119**, 1562–1573 (2006).

METHODS

Preprocessing

A detailed description of preprocessing of the neural data can be found in [24]. In short, data preprocessing included montaging to a common average reference, noise removal, extraction of the high gamma envelope (70 – 150 Hz) using the Hilbert transform. The high-gamma envelope is thought to reflect neuronal population activity[44,45]. Finally, neural responses were down sampled to 100 Hz and z-scored across single speaker blocks and across multi-talker blocks (i.e. calculated over both male and female trials, but separately for single- and multi-source blocks).

Speech responsive electrodes

To assess which electrodes exhibited a robust response to speech streams, we computed for each electrode the mean baseline response as the average of the high-gamma envelope during 0.5 seconds preceding stimulus onset, and the mean speech onset response as the average of the high-gamma envelope in the 0.5 seconds following stimulus onset. To test for a statistically significant auditory response, we performed a paired samples t-test for each electrode and applied FDR correction across electrodes to correct for multiple comparisons. Only electrodes that exhibited a robust auditory response at $q < 0.05$ were included in the remainder of the analysis.

Estimating spectrotemporal receptive fields (STRFs) and response latency

First, we computed a cortical spectrogram representation of each sound scene using a model of early cochlear processing and mid-brain auditory processing (NSL toolbox[1]). We modeled cochlear processing using a filter bank of 128 constant-Q filters that were spaced equally on a logarithmic axis ranging from center frequency (CF) = 270 Hz to CF = 7,246 Hz. Next, we modeled auditory midbrain processing by taking the derivative along the frequency axis, performing half-wave rectification and applying short-term temporal integration[46,47]. This approach accounted for the enhanced frequency selectivity as a consequence of lateral inhibition, as well as reduced phase locking, observed after midbrain processing. Cortical spectrograms were computed based on monaural stimulus waveforms (i.e. independent of sound location). The resulting spectrograms had a sampling frequency of 100 Hz and were down sampled to 50 channels to reduce the number of parameters.

We then estimated the spectrotemporal receptive field (STRF) by linearly mapping the cortical spectrogram to the evoked response using the STRFlab MATLAB Toolbox[48] (http://strflab.berkeley.edu). For each electrode, we used the past 300 ms of a stimulus to predict the neural response at every time point using normalized reverse correlation. To prevent overfitting, we used a five-fold cross-validation

[1] Available from http://nsl.isr.umd.edu/index.html

682 procedure. We optimized sparsity and regularization parameters by maximizing the correlation between
683 actual and predicted responses. Using the resulting STRFs, we defined the response latency for each
684 electrode as the time point corresponding to the peak energy in the STRF.

685 **Decoding a single talker's voice and location features**

686 We trained a four-class classifier on population neural response patterns to jointly decode a talker's
687 voice and location features. The four classes corresponded to 'male talker, left', 'male talker, right',
688 'female talker, left' and 'female talker, right'. We used frame-by-frame, regularized least-squares (RLS)
689 classification[22,33] which produced for each time frame a linear weighted sum of the population of neural
690 responses for each class[22]. We trained and tested classifiers on the sustained responses only (i.e.,
691 excluding response onset effects from 0 to 500 ms). The class with the highest average classifier output
692 over all frames in the trial was taken as the predicted class.

693 We trained the classifier in a leave-two-trials-out cross-validation procedure on the single-talker data
694 (corresponding to 25 folds). We computed classification accuracy as the average accuracy across the 25
695 folds. Further, to evaluate the statistical significance of classification accuracies, we performed a
696 permutation analysis in which we randomly permuted the class labels and repeated the complete 25-
697 fold cross-validation procedure. We iterated this process 2,000 times to create a null distribution of
698 classification accuracy. Next, we tested whether the observed classification accuracy exceeds the 95th
699 percentile of the null distribution of permuted accuracies (one sided test). We computed p as the
700 proportion of permuted accuracies that was equal to or larger than the observed accuracy.

701 Finally, we calculated marginal accuracies for the voice and location feature dimensions by labelling
702 accuracy based on a single feature dimension only, ignoring the other feature dimension. For example,
703 to quantify the marginal accuracy for voice features, we calculated the percentage of the trials for which
704 the correct voice class was predicted (i.e. female or male talker), ignoring the predicted location class
705 (i.e. left or right). We computed the marginal accuracy also as the average across the 25 folds and used
706 the permutation procedure described above to assess the statistical significance of the marginal
707 accuracies.

708 **Attentional-driven response gains in multi-talker scenes**

709 For each electrode, we quantified the strength and direction of attentional modulations of cortical
710 responses in the multi-source scenes evoked either by attending to a talker's voice features or by
711 attending to a talker's location features using Cohen's d. That is, similar to the quantification of single-
712 talker feature sensitivity described above, we computed the mean response for each trial in the multi-
713 source condition as the mean from 0.5 s post sound onset to 1.5 s post sound onset. Then, to test for
714 attention-driven response gains for a talker's voice features, we computed the effect size for the

715 difference between the mean responses to all 'attend male' and 'attend female' trials, irrespective of the
716 attended location of the trials (n = 50 each). To test for attention-driven response gains for a talker's
717 location features, we computed the effect size for the difference between the mean responses to all
718 'attend left' and 'attend right' trials, irrespective of the attended talker of the trials (n = 50 each).

719 **Decoding an attended talker's voice and location features in multi-talker scenes**

720 To decode an attended talker's voice and location features in spatial multi-talker scenes, we trained the
721 four-class classifier on the multi-talker data using a similar procedure as described above. In multi-talker
722 scenes, class labels consisted of 'attended male talker, left', 'attended male talker, right', attended female
723 talker, left' and 'attended female talker, right'. We also assessed statistical significance using a
724 permutation procedure similar to the permutation procedure for single talker data.

725 **Quantifying temporal coherence**

726 We assessed temporal coherence in slow fluctuations in stimulus evoked responses between pairs of
727 voice sensitive and location sensitive sites. This analysis was performed on a within subject level. Five
728 subjects contained multiple voice-location pairs of and were therefore included in the analysis. Because
729 the high-gamma envelope is considered a signature of neural population responses[44,45], we computed
730 temporal coherence on the high-gamma envelope. Further, we quantified temporal coherence using the
731 coherency coefficient, which is the mathematical equivalent in the frequency domain of the cross-
732 correlation function in the time domain[38]. Specifically, the coherence coefficient is the normalized
733 average cross-power spectral density between signals x and y across trials at frequency ω computed
734 as[38]:

735 $$coh_{xy}(\omega) = \frac{\left|\frac{1}{n}\sum_{k=1}^{n} A_x(\omega,k)A_y(\omega,k)e^{i(\varphi_x(\omega,k)-\varphi_y)}\right|}{\sqrt{\left(\left(\frac{1}{n}\sum_{k=1}^{n} A_x^2(\omega,k)\right)\left(\frac{1}{n}\sum_{k=1}^{n} A_y^2(\omega,k)\right)\right)}}$$

736 Here, we computed broadband temporal coherence over a frequency range of 2 – 22 Hz to map the
737 development of attentional enhancement of temporal coherence over time, correspond to the range of
738 slow fluctuations in which temporal coherence for feature binding is hypothesized to occur (i.e. 50 ms
739 to 500 ms[15]). Furthermore, we computed narrowband temporal coherency for eight frequency bands
740 with center frequencies 3, 6, 9, 12, 15, 18, and 21 Hz (bandwidth = 3 Hz) to examine the effect of attention
741 on temporal coherence for specific frequencies.

742

ACKNOWLEDGEMENTS

We thank Menoua Keshishian for sharing his expertise on STRF analysis. This study was supported by National Institute on Deafness and other Communication Disorders grant R01DC014279 (NM) and grant R01DC018805 (NM). This project has also received funding from the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie grant agreement No 898134 (KH) and from the NWO Talent Program under the Veni grant agreement VI.Veni.202.184 (KH).