Hierarchical cortical networks of “voice patches” for processing voices in human brain

Yang Zhang[a,b,c,1], Yue Ding[a,b,d], Juan Huang[c], Wenjing Zhou[e], Zhipei Ling[f], Bo Hong[a,b,1], and Xiaoqin Wang[a,b,c,1]

[a]Tsinghua Laboratory of Brain and Intelligence (THBI), Tsinghua University, Beijing 100084, China; [b]Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China; [c]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [d]Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China; [e]Department of Epilepsy Center, Tsinghua University Yuquan Hospital, Beijing 100040, China; and [f]Department of Neurosurgery, General Hospital of People's Liberation Army, Beijing 100853, China

Edited by Peter Strick, Neurobiology, University of Pittsburgh, Pittsburgh, PA; received July 27, 2021; accepted November 11, 2021

Humans have an extraordinary ability to recognize and differentiate voices. It is yet unclear whether voices are uniquely processed in the human brain. To explore the underlying neural mechanisms of voice processing, we recorded electrocorticographic signals from intracranial electrodes in epilepsy patients while they listened to six different categories of voice and nonvoice sounds. Subregions in the temporal lobe exhibited preferences for distinct voice stimuli, which were defined as “voice patches.” Latency analyses suggested a dual hierarchical organization of the voice patches. We also found that voice patches were functionally connected under both task-engaged and resting states. Furthermore, the left motor areas were coactivated and correlated with the temporal voice patches during the sound-listening task. Taken together, this work reveals hierarchical cortical networks in the human brain for processing human voices.

voice patch | human brain | ECoG | dual pathway

The ability to recognize and differentiate sound categories is crucial to survival for many species, including human beings. In both humans and nonhuman primates, sound recognition is thought to be accomplished primarily in the ventral auditory pathway (1–6), which includes structures in the anterior and middle portions of the temporal lobe. The major question regarding the underlying mechanism of sound recognition is whether the representation of different categories of sounds is distributed along the entire ventral auditory stream or is localized in distinct regions.

For humans and many other animal species, the most important category of sounds is their species-specific voices or vocalizations. The human voice contains not only speech information but also a wealth of information about the speaker’s identity and emotional state. Recognizing this information carried by the human voice is important for our social interactions. Human functional imaging study (7–9) has demonstrated the existence of voice-specific cortical regions, which are located on the lateral superior temporal gyrus (STG) and in the upper bank of superior temporal sulcus (STS). These regions have been shown to prefer human voices and animal vocalizations over acoustic controls and natural sounds (7, 10–12). Studies with nonhuman primates also have demonstrated the existence of vocalization-specific regions in macaques (13) and marmosets (14), which are located on the rostral portion of auditory cortex. These results reveal that the voice/vocalization-specific regions are evolutionarily conserved in primates.

Voice and vocalization can be considered as “auditory face” due to similar types of socially relevant information carried by faces and voices/vocalizations. In visual cortex, the face patch system was found to be specialized for processing faces in humans and nonhuman primates (15–18). This face patch system consists of a series of discrete and interconnected cortical areas that are selective to faces. Analogous to the face patch system of visual cortex, the notion of a “voice patch system” has been established (9, 19). However, the evidence to support this notion is fragmentary. In humans, clustering analysis on voice sensitivity peaks of functional MRI (fMRI) signals across subjects suggests three “voice patches” (anterior, middle, and posterior STG) along the STG bilaterally (9), suggesting the existence of a voice patch system in the human brain. However, these data raise important questions. What are the functional roles of each voice patch, and what are the relations between voice patches? The fMRI methodology used in previous studies on voice- and vocalization-specific areas lacks the temporal precision to infer the dynamics between these cortical areas involved in processing voices and vocalizations.

To address these challenges, we recorded electrocorticographic signals (ECoG) from epilepsy patients while they were presented with six different categories of voice and nonvoice sounds. We have two goals in this study. Our first goal was to identify if there exist voice patches that are selective for voice over nonvoice sounds, and if so, what are the response properties of each voice patch. The second goal was to investigate the connectivity between voice patches. We identified three voice patches along the STG in each hemisphere. The voice patches were hierarchically organized along a dual pathway and functionally connected under both task-engaged and resting

Significance

The human voice contains information about a speaker's identity and emotional state. How the brain processes the human voice remains largely unknown. We recorded electrocorticographic signals from intracranial electrodes implanted in epilepsy patients while they listened to different categories of voice and nonvoice sounds. We identified several spatially distinct “voice patches” in the temporal lobe that exhibited preferences for human voices. Further analyses suggested that the voice patches are functionally connected, and form a dual-directional hierarchical network for voice processing. This study provides clear evidence to demonstrate the existence of an interconnected “voice patch system” in the human brain, which is analogous to the face patch system of primate visual cortex, suggesting similar cortical architectures for processing faces and voices.

Author contributions: Y.Z., J.H., B.H., and X.W. designed research; Y.Z. and Y.D. performed research; W.Z. and Z.L. performed the neurosurgeries on the epilepsy patients; Y.Z. analyzed data; and Y.Z., J.H., B.H., and X.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yzhan142@jhu.edu, hongbo@tsinghua.edu.cn, or xiaoqin.wang@jhu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2113887118/-/DCSupplemental.

Published December 20, 2021.

states. In addition, the left motor areas were also found to be involved in human voice processing.

Results

Diverse Response Patterns of Individual Electrodes to Six Categories of Sounds. Five subjects with intracranial surface electrodes covering temporal, frontal, and parietal lobes were included in this study (*SI Appendix*, Fig. S1). All subjects speak Chinese as their native language and have no English background. ECoG signals were recorded while six different categories of sounds were presented to them (sound–listening tasks). The six categories of sounds (*SI Appendix*, Fig. S2) are Chinese speech (CS, voice sound), English speech (ES, voice sound), Nonspeech voice (NSV, voice sound), animal vocalizations (AV, voice sound), natural sounds (NS, nonvoice sound), and scrambled sounds (SS, nonvoice sound). Fig. 1*A* shows electrode positions (green dots) on the reconstructed brain surface in two subjects (subject 4: right hemisphere; subject 5: left hemisphere). Electrodes with significant responses (compared to baseline, $P < 0.05$) are labeled with red symbols (shapes of symbols indicating different types of response patterns as shown in Fig. 1*C*). Fig. 1*B* shows the response recorded from a representative electrode plotted in the spectral–temporal domain (*Upper*). This electrode has significant response to the Chinese word “you” in high gamma (HG) frequency range (70 to 140 Hz). The energy level across the HG frequency range is averaged and plotted as Z-score waveform (Fig. 1 *B*, *Lower*). For each electrode, we considered it as a responsive electrode if its peak Z-score is higher than 2.

We then calculated the Z-score waveforms to the six categories of sound stimuli (arranged in an order from CS, ES, NSV, AV, NS, to SS; each row represents a stimulus as shown in Fig. 1*C*) of all responsive electrodes from the five subjects. For each electrode, the Z-score waveform was normalized to the maximum Z-score across all tested stimuli. The normalized Z-score waveforms of all responsive electrodes from all subjects were divided into five types of response patterns based on their sound selectivity. Fig. 1*C* shows the five types of response patterns from five representative electrodes (arranged in five columns: e1 to e5). “e1-type” represents the electrodes that responded to all categories of stimuli (no selectivity), “e2-type” represents the electrodes that responded to voices and vocalizations (CS, ES, NSV, and AV) but not to natural and scrambled sounds (NS and SS), “e3-type” represents the electrodes that only responded to human voices (CS, ES, and NSV), “e4-type” represents the electrodes that only responded to human voiced speech (VS [CS and ES]), and “e5-type” represents the electrodes that showed the highest selectivity—they only responded to CS (native language for all subjects), respectively. For all electrodes tested, we observed electrode selectivity for voice or vocalization stimuli (CS, ES, NSV, and AV) but not for nonvoice stimuli (NS and SS) in our experiments. Some electrodes showed selectivity to only a subset of voice stimuli (e.g., e4-type and e5-type, Fig. 1*C*).

Voice Patches on Temporal Lobe. Fig. 1*A* shows the spatial locations of the electrodes in two subjects with the five types of response patterns depicted in Fig. 1*C*. Electrodes (green dots)

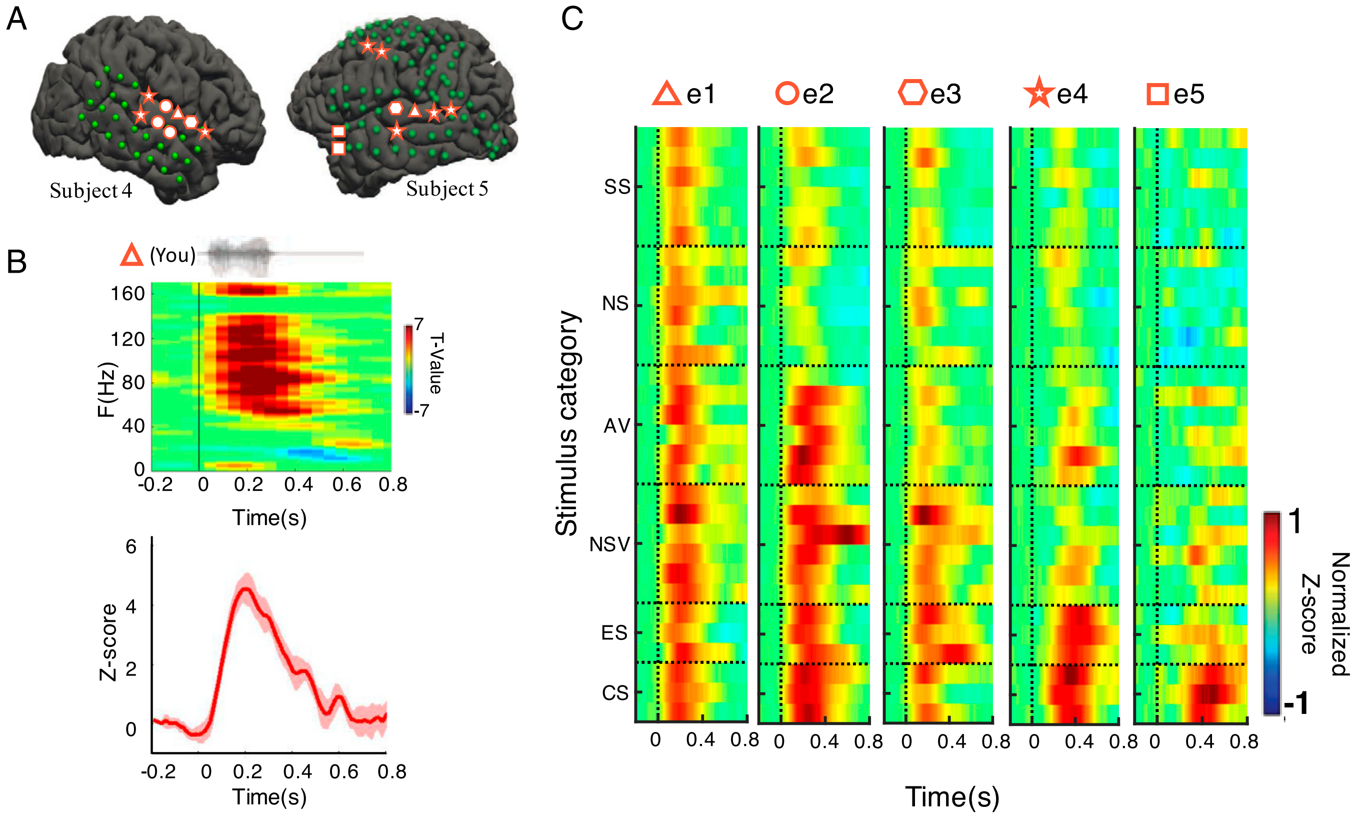

Fig. 1. Response patterns of individual electrodes to six categories of sounds. (*A*) MRI surface reconstruction of two subjects (subject 4: right hemisphere; subject 5: left hemisphere). Green dots represent all electrodes implanted in the subjects. The red symbols overlaid on the electrodes represent different response patterns shown in *C*. (*B*) Example stimulus (Chinese word “You/有/”), spectral–temporal (*Upper*), and Z-score (mean ± SEM, *Lower*) responses of one representative electrode from subject 4 (red triangle in subject 4) in response to the example stimulus. (*C*) Normalized Z-score responses from five representative elevctrodes (e1 through e5, each column represents one electrode) to all stimuli (30 stimuli, each row represents one stimulus). These responses represent the five response patterns found in all electrodes. Each response pattern is marked with a red symbol.

with significant responses (marked with red symbols) to the tested sound stimuli were mainly found in the posterior STG, middle STG, anterior STG, and motor areas in these two representative subjects (Fig. 1*A*). To quantify the category selectivity of cortical responses, we computed a category selectivity index (CSI, see *Materials and Methods*) for each electrode. CSI measures the distance between responses to selective categories (the mean response to this category of sounds is higher than the mean response across all stimuli) and to nonselective categories (the mean response to this category of sounds is lower than the mean response across all stimuli). Fig. 2*A* shows the CSI of all electrodes from both hemispheres (n = 384, black dots) recorded in all subjects. Electrodes with high CSI values are clustered in the temporal lobe and motor areas in both hemispheres, suggesting that these areas have high sound category selectivity. The electrodes with high CSI values on the temporal lobes of both hemispheres (motor areas will be discussed later) can be grouped into three patches (Fig. 2*A*), which were referred to as posterior temporal patch (PT), middle temporal patch (MT), and anterior temporal patch (AT). The spatial locations of electrodes on the temporal lobes with CSI values greater than 0.33 (corresponding to a 2:1 ratio of selective-to-nonselective category responses) are shown in Fig. 2*B*. Electrodes from different patches (PT, MT, and AT) were labeled in different colors (PT: orange; MT: blue; AT: purple), and the electrodes were numbered from caudal to rostral. In total, we identified six electrodes in the PT, six electrodes in the MT, and three electrodes in the AT of the left hemisphere and two electrodes in the PT, six electrodes in the MT, and three electrodes in the AT of the right hemisphere.

We then calculated the response amplitude of each patch to the six categories of sounds by averaging the maximum values of the Z-score waveforms (maximum Z-score) across all electrodes in each patch (Fig. 2*C*). The PT in both hemispheres showed significant responses to human VS (CS and ES, Fig. 2 *C*, *Upper Left* and *Upper Right*). The MT in both hemispheres showed more significant responses to voices and vocalizations (CS, ES, NSV, and AV, Fig. 2 *C*, *Middle Left* and *Middle Right*) than to nonvoices (NS and SS). However, we observed differences in the responses of the AT between hemispheres. The AT in the left hemisphere showed significant responses only to CS (Fig. 2 *C*, *Lower Left*), while the AT in the right hemisphere showed significant responses to human VS (CS and ES, Fig. 2 *C*, *Lower Right*). We defined a selectivity index as $\mathrm{SI}_{(\mathrm{CS\ versus\ ES})}$, $\mathrm{SI}_{(\mathrm{CS\ versus\ ES})} = (R_{\mathrm{CS}} - R_{\mathrm{ES}})/(R_{\mathrm{CS}} + R_{\mathrm{ES}})$ to quantify the distance between responses to CS and ES (see *Materials and Methods*). When $\mathrm{SI}_{(\mathrm{CS\ versus\ ES})}$ is positive, it indicates a stronger response to CS than ES. $\mathrm{SI}_{(\mathrm{CS\ versus\ ES})}$ is negative if a stronger response to ES than CS is found. *SI Appendix*, Fig. S3*A* shows the $\mathrm{SI}_{(\mathrm{CS\ versus\ ES})}$ across the whole brain. We found that AT in the left hemisphere has the highest selectivity to CS versus ES. Distributions of $\mathrm{SI}_{(\mathrm{CS\ versus\ ES})}$ across all electrodes showed significant differences between the left and right hemispheres, specifically that the left hemisphere was more selective to CS than the right hemisphere (*SI Appendix*, Fig. S3*B*, $P < 0.01$, two-sample Student's t test). These results provided evidence for left lateralization of native language processing.

Overall, we found that all patches have significantly stronger responses to voice/vocalization stimuli or to a subset of voice stimuli than nonvoice stimuli (Fig. 2*C*, $P < 0.05$, rank sum test). Considering the voice/vocalization selectivity of the patches and their disconnected spatial locations, we referred to these patches as voice patches in the human brain.

Response Properties of Voice Patches. Given the voice/vocalization selectivity of each voice patch, we sought to further characterize the response properties of each voice patch by investigating the voice selectivity and the response latency of each electrode within the voice patch. We constructed a matrix containing the response patterns for all electrodes across all voice patches in both hemispheres (Fig. 3*A*). In this matrix, each column corresponds to a single electrode (the order of the electrodes is derived from Fig. 2*B*), and each row corresponds to a single sound stimulus. The sound stimuli are grouped into six categories showed in the y-axis. Electrodes from different voice patches are grouped in

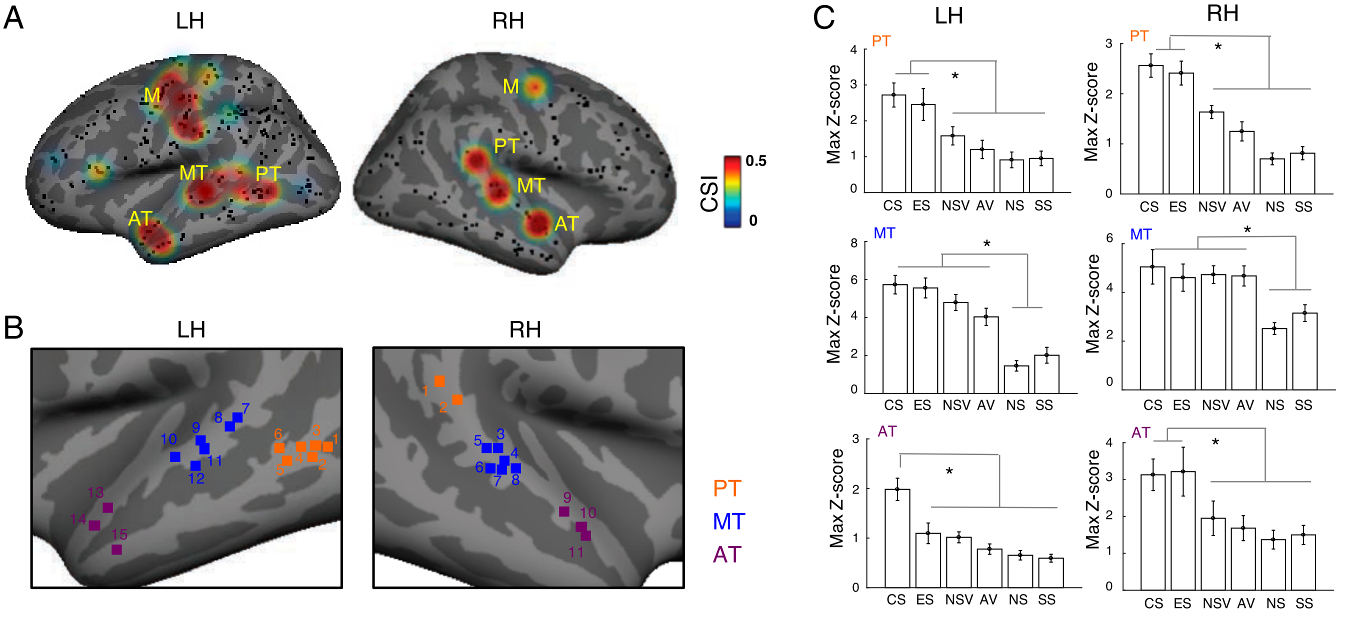

Fig. 2. Electrodes with voice selectivity are grouped into several voice patches in temporal areas. (*A*) Distribution of CSI across all electrodes (black dots represent all implanted electrodes in all subjects) in the left and right hemispheres on the inflated average brain. Three patches were found in temporal lobes of each hemisphere (M: motor; motor areas will be discussed later). (*B*) Precise locations of all voice-selective electrodes (CSI > 0.33) on the inflated average brain (electrodes in the motor areas were not shown, different colors of the electrodes indicate different patches, and the electrode number was ordered from caudal to rostral). (*C*) Maximum Z-score responses (mean ± SEM) averaged across all electrodes from each voice patch (*$P < 0.05$, rank sum test).

https://doi.org/10.1073/pnas.2113887118

rectangles with different colors shown in the *x*-axis, and voice patches are separated by the red dashed lines. Electrodes in the MT of both hemispheres responded to more categories of sounds than electrodes in the AT and the PT. We computed the number of selective categories for each electrode and averaged across all electrodes in each voice patch (Fig. 3*B*). The MT had a significantly higher number of selective categories than those of the PT and the AT in both hemispheres.

We also computed response latencies for all electrodes shown in Fig. 3*A* (Fig. 3*C*, electrodes are in the same order as Fig. 3*A*). Only CS responses were included in the latency analyses since all electrodes responded to this sound category (Fig. 1*C*). Latency was defined as the time point relative to sound onset when HG power first exceeds the 95% confidence interval (CI) of the baseline mean of each responsive trial (trial with peak Z-score higher than 2) as described in a previous study (20). Fig. 3*C* shows the latencies of all electrodes in the same order as in Fig. 3*A* in both hemispheres. Electrodes in the MT showed the shortest latencies, whereas the electrodes in the PT and the AT had longer latencies. By averaging the latencies across all electrodes in each voice patch, we showed the latencies of the MT are significantly shorter than those of the PT and the AT (Fig. 3*D*). These findings suggest that the MT is activated by sound stimuli prior to the PT and the AT.

To compare the response amplitudes of voice patches, we calculated the maximum Z-score of each electrode in response to CS stimuli and averaged across all electrodes in each voice patch (Fig. 3*E*). The MT showed a significantly higher response amplitude than the PT and the AT in both hemispheres ($P < 0.05$, rank sum test). We also computed a sparseness value for each electrode in response to CS stimuli to represent response stability. Sparseness has a value between 0 and 1, with lower sparseness indicating higher response stability. High response stability suggests that the responses more faithfully follow the external stimuli. Therefore, areas with higher response stability are likely to be at a lower processing level (21). Fig. 3*F* shows the sparseness value of each voice patch by averaging the sparseness values of all electrodes in that voice patch. The MT has significant lower sparseness values than the PT and the AT in both hemispheres ($P < 0.05$, rank sum test).

Taken together, these results suggest that the MTs of both hemispheres function as initial hubs for processing voices given their shortest response latencies, responses to broadest categories of sounds, highest response amplitude, and lowest

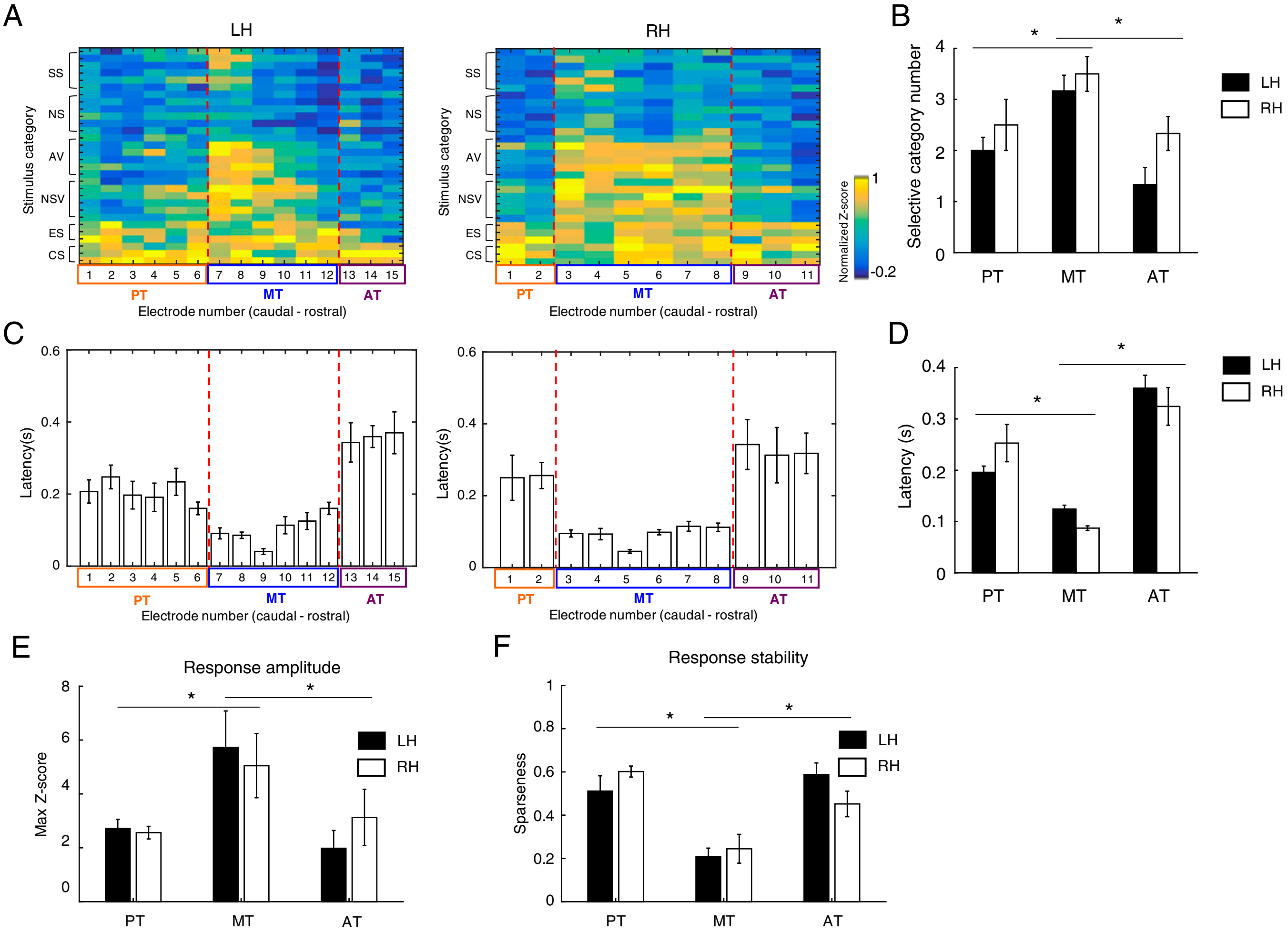

Fig. 3. Properties of voice patches. (*A*) Response patterns of all voice selective electrodes (CSI > 0.33) in the left (LH) and right (RH) hemispheres, electrodes are listed in an order shown in Fig. 2*B* (from caudal to rostral), electrodes inside each rectangle showed in the *x*-axis are from the same voice patch (orange rectangle: PT; blue rectangle: MT; purple rectangle: AT). Different voice patches are separated by the red dash lines. (*B*) Number of selective categories averaged across all electrodes from each voice patch (*$P < 0.05$, rank sum test). (*C*) Latencies of all voice selective electrodes, electrodes are listed in the same order as *A*. (*D*) Latencies averaged across all electrodes from each voice patch (*$P < 0.05$, rank sum test). (*E*) HG response amplitude (mean ± standard deviation [STD]) averaged over all electrodes from each voice patch in the left and right hemispheres (*$P < 0.05$, rank sum test). (*F*) Sparseness (mean ± STD) averaged over all electrodes from each voice patch in the left and right hemispheres (*$P < 0.05$, rank sum test).

sparseness values. The AT and the PT are downstream of the information flow from the MT to further process a selective subset of voices.

Connectivity of Voice Patches under Task-Engaged and Resting States. We next investigated the connectivity of voice patches both under task-engaged and resting states. In the task-engaged state, we compared the similarity of responses of each voice patch by correlating the maximum Z-score across all trials under the CS condition (22). Only electrodes from the same subject and with their responses recorded simultaneously were included in the analyses. Fig. 4*A* shows four example electrodes (electrodes 1 through 4). Electrode 1 was located in the AT of the left hemisphere and was chosen as the example reference electrode, electrode 2 was located in the AT of the left hemisphere and was within the same voice patch as electrode 1 (within-patch pair), electrode 3 was located in the PT of the left hemisphere and was in a voice patch different from electrode 1 (between-patch pair), and electrode 4 was located outside of any voice patches (outside-patch pair). Fig. 4*B* shows the maximum Z-score responses across all trials under CS condition for these four electrodes. We then computed the Pearson's correlations between the reference electrode 1 and the other three electrodes (Fig. 4 *C*, *Left*), and the significance was confirmed by permutation tests (*SI Appendix*, Fig. S4, $P < 0.05$ was chosen as the criterion). We observed significant correlations between electrodes 1 and 2 (within-patch pair) and between electrodes 1 and 3 (between-patch pair), and the correlation between the within-patch pair was higher than that of the between-patch pair. No significant correlation was found between electrodes 1 and 4 (outside-patch pair).

Functional imaging studies (23, 24) and ECoG studies (22, 25) all indicate that the low frequency fluctuations (<1 Hz) of both BOLD (blood oxygen level–dependent) and ECoG signals under the resting state can be used to determine the intrinsic functional connectivity of different brain areas. In this study, to determine the intrinsic functional connectivity between voice patches, we extracted the slow fluctuations of the HG band envelop of resting state ECoG signals. *SI Appendix*, Fig. S5*A* shows the resting state HG envelop of the four example electrodes (electrodes 1 through 4 in Fig. 4*A*). Pearson's correlations were then calculated between the reference electrode 1 and the other three electrodes (Fig. 4 *C*, *Right*), and the significance was confirmed by permutation tests (*SI Appendix*, Fig. S5 *B*–*D*, $P < 0.05$ was chosen as the criterion). Similar to that seen under the task-engaged state, we also observed significant correlations under the resting state between electrodes 1 and 2 (within-patch pair) and between electrodes 1 and 3 (between-patch pair), and the correlation between the within-patch pair was higher than that of the between-patch pair. Electrodes 1 and 4 (outside-patch pair) also showed a significant correlation under resting state (Fig. 4 *C*, *Right*); however, its value is much lower than that of either the within-patch pair or the between-patch pair.

To generalize this observation, we performed the same task-engaged state trial-based maximum Z-score correlation and resting state HG envelop correlation across all unique pairs of all electrodes in each subject, including all within-patch pairs,

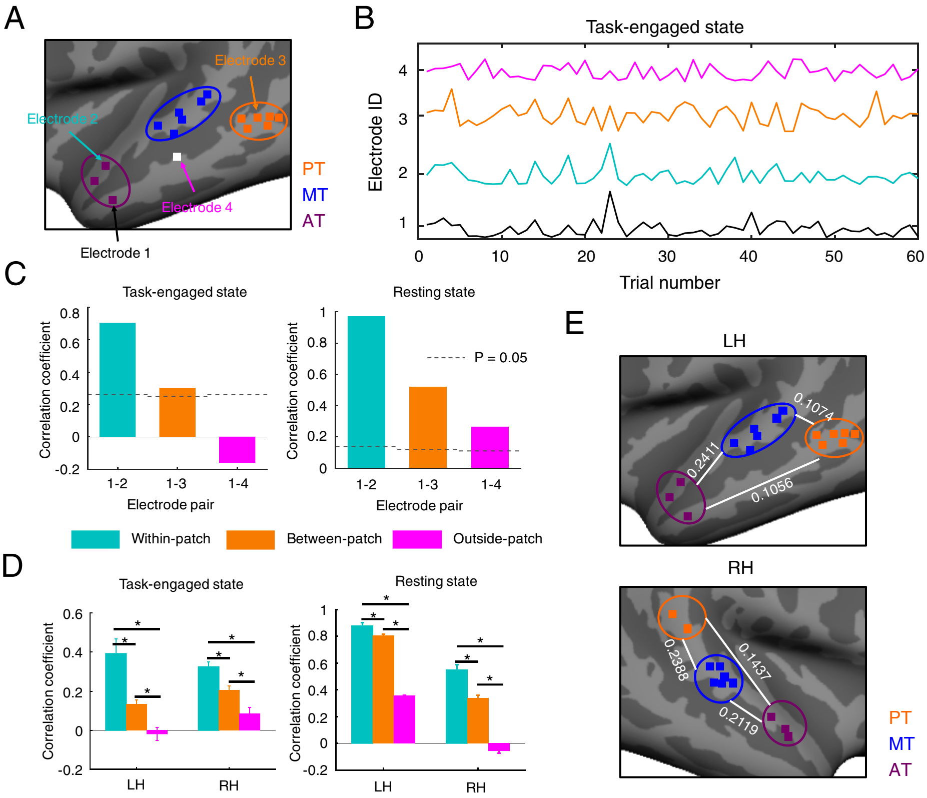

Fig. 4. Connectivity of voice patches. (*A*) Four example electrodes are chosen (electrode 1: reference electrode; electrode 2: within-patch electrode of electrode 1; electrode 3: between-patch electrode of electrode 1; electrode 4: outside-patch electrode of electrode 1). (*B*) Maximum Z-score responses of the four example electrodes across all trials under CS condition. (*C*) Correlations between the reference electrode (electrode 1) and the other three electrodes under task-engaged and resting states (significance levels were determined by permutation tests). (*D*) Comparisons of correlations from all within-patch electrode pairs, all between-patch electrode pairs, and all outside-patch electrode pairs in both hemispheres under task-engaged and resting states (*$P < 0.05$, rank sum test). (*E*) Diagrams of connectivity patterns between all voice patches in the left (upper) and right (lower) hemispheres under task-engaged state (value: mean correlation coefficient).

all between-patch pairs, and all outside-patch pairs. We constructed a task-engaged state correlation matrix and a resting state correlation matrix for each subject. *SI Appendix*, Fig. S6 *A* and *B* show the correlation matrices of all electrodes from all voice patches in subject 5 (electrodes implanted in the left hemisphere) under task-engaged and resting states, respectively. There is a significant positive correlation between connectivity matrices across states in subject 5 (*SI Appendix*, Fig. S6*C*, $P = 0.0025$, $r = 0.4878$). *SI Appendix*, Fig. S6 *D–E* show the correlation matrices of subject 4 (electrodes implanted in the right hemisphere) under task-engaged and resting states, respectively. We also observed a significant positive correlation between connectivity matrices across states in subject 4 (*SI Appendix*, Fig. S6*F*, $P = 0.0453$, $r = 0.3357$). By averaging correlations across all within-patch electrode pairs, between-patch electrode pairs, and outside-patch electrode pairs from all subjects, we found that the correlations for within-patch and between-patch electrode pairs were significantly higher than those for outside-patch electrode pairs and that the correlations for within-patch electrode pairs were significantly higher than those for between-patch electrode pairs in both task-engaged and resting states (Fig. 4*D*, $P < 0.05$, rank sum test). These results suggest that electrodes within a voice patch and between voice patches are highly connected and that these connectivity patterns are similar in both task-engaged and resting states. To compare the connectivity strength between different voice patches, we averaged the correlations for each unique voice patch pair across all subjects in both task-engaged and resting states (*SI Appendix*, Fig. S7). No significant differences were found between the voice patch pairs in either task-engaged or resting states ($P > 0.05$, rank sum test), suggesting that all voice patches are connected at a similar intensity level. Fig. 4*E* shows the connectivity diagrams between all voice patch pairs of both hemispheres.

Alternative frequency bands of the resting ECoG were also used to infer the intrinsic functional connectivity between voice patches. We calculated the correlation patterns across all unique pairs of all electrodes in each subject for the delta (1 to 3 Hz), theta (4 to 7 Hz), alpha (8 to 12 Hz), beta (12 to 20 Hz), gamma (20 to 40 Hz), and HG (70 to 140 Hz) ranges (*SI Appendix*, Fig. S8). The analyses were performed separately for within-patch electrode pairs and between-patch electrode pairs. Results from example electrode pairs (*SI Appendix*, Fig. S8 *A* and *B*) and population analyses (*SI Appendix*, Fig. S8 *C* and *D*) showed that significant higher correlations were found by using slow fluctuations of HG-filtered signals, which suggest that using HG-filtered signals is the best option to infer connectivity between electrodes. Similar results were also shown in previous studies (24, 26).

Involvements of Motor Areas in Voice Processing. Fig. 2*A* shows the CSI distribution for both hemispheres from all subjects. High CSI values were also found in motor areas. We calculated the response amplitude of motor areas to the six categories of sounds by averaging the maximum Z-scores across all electrodes with CSI > 0.33 in motor areas of both hemispheres (Fig. 5 *A*, *Left hemisphere*: $n = 5$; *Right hemisphere*: $n = 3$). Motor areas in the left hemisphere showed significantly higher responses to human VS (CS and ES, Fig. 5 *A*, *Left*) than to other sounds, while motor areas in the right hemisphere showed no significant responses (Fig. 5 *A*, *Right*). To further understand the roles of motor areas in the sound–listening tasks, we used a seed-based approach to access the correlations between motor areas and temporal voice patches. Fig. 5*B* shows the locations of example seed electrodes in both left and right motor areas. Correlations between the example seed electrodes and all electrodes in the temporal voice patches were calculated for both hemispheres across all trials under each sound category condition. Notably, only electrodes from the same subject and with their responses recorded simultaneously were included in the analyses. Strong correlations were observed only in the PT and the AT of the left hemisphere under CS and ES conditions (Fig. 5*C*). We then calculated the correlations between all seed electrodes in the motor areas and all electrodes in the temporal voice patches in response to the six categories of sounds. Fig. 5*D* shows the mean correlation coefficients between seed electrodes in the motor areas and electrodes in the temporal voice patches. Significant correlations were only observed in the PT and the AT in the left hemisphere under CS and ES conditions (Fig. 5*D*, $P < 0.05$, significance was confirmed by permutation test). These results suggest that motor areas in the left hemisphere are significantly correlated with left higher-level temporal voice patches (PT and AT) in speech processing during listening tasks, while motor areas in the right hemisphere showed no such response properties.

Discussion

By analyzing the cortical neural responses to six different categories of sounds, we identified five different selectivity patterns of individual electrodes in the human brain (Fig. 1*C*). Combining voice selectivity with response latencies and spatial locations of all electrodes from all patients in both hemispheres led to the localization of voice patches (three voice patches in each hemisphere, Fig. 2). The analyses of response latencies and properties (Fig. 3) of the voice patches suggested a dynamic dual information flow in both hemispheres in which MT voice patches are the initial hubs. Further analyses suggested voice patches are functionally connected under both task-engaged and resting states (Fig. 4). In addition, the left motor areas were found to be involved in the sound–listening tasks for processing speech sounds (Fig. 5). These findings provide insights into how human voices are processed in the human brain and how the voice patches are interconnected to support voice perception, suggesting similar cortical architectures for processing faces and voices.

A Network of Voice Patches for Processing Human Voices. Previous studies have shown that voice-specific areas are located in the middle lateral STG and upper bank of STS in the human brain and respond significantly higher to voices than to nonvoice control sounds (7–9). In this study, MT voice patches in both hemispheres showed stronger responses (Fig. 2*C*) to voices (CS, ES, NSV, and AV) than to nonvoices (NS and SS), consistent with the findings of voice-specific areas identified using functional imaging methods (7). Therefore, our work validates the existence of these voice-specific areas in the MT areas (left and right) in the human brain with more neurophysiological data. In addition to the MT voice patches (left and right), we also found four other voice patches located in the PT and the AT (left and right), suggesting that voice processing in the human brain is dependent on a network that consists of multiple cortical areas. However, the grid recording used in the present study can only capture responses on the STG, and thus, it remains unclear whether there are any other voice-specific clusters in the STS.

Previous studies have also provided evidence regarding the functions of these specific brain areas. Studies showed that PT areas are involved in the encoding of phonetic features (27), consistent with our findings that PT voice patches (left and right) show selective responses to human VS (CS and ES), as they are the only stimuli that contain phonetic structures used in the study. AT areas are found to be involved in higher cognitive functions. For example, studies by Belin and Zatorre using an fMRI adaptation paradigm (28) showed that the activity of right anterior STG was significantly reduced when syllables

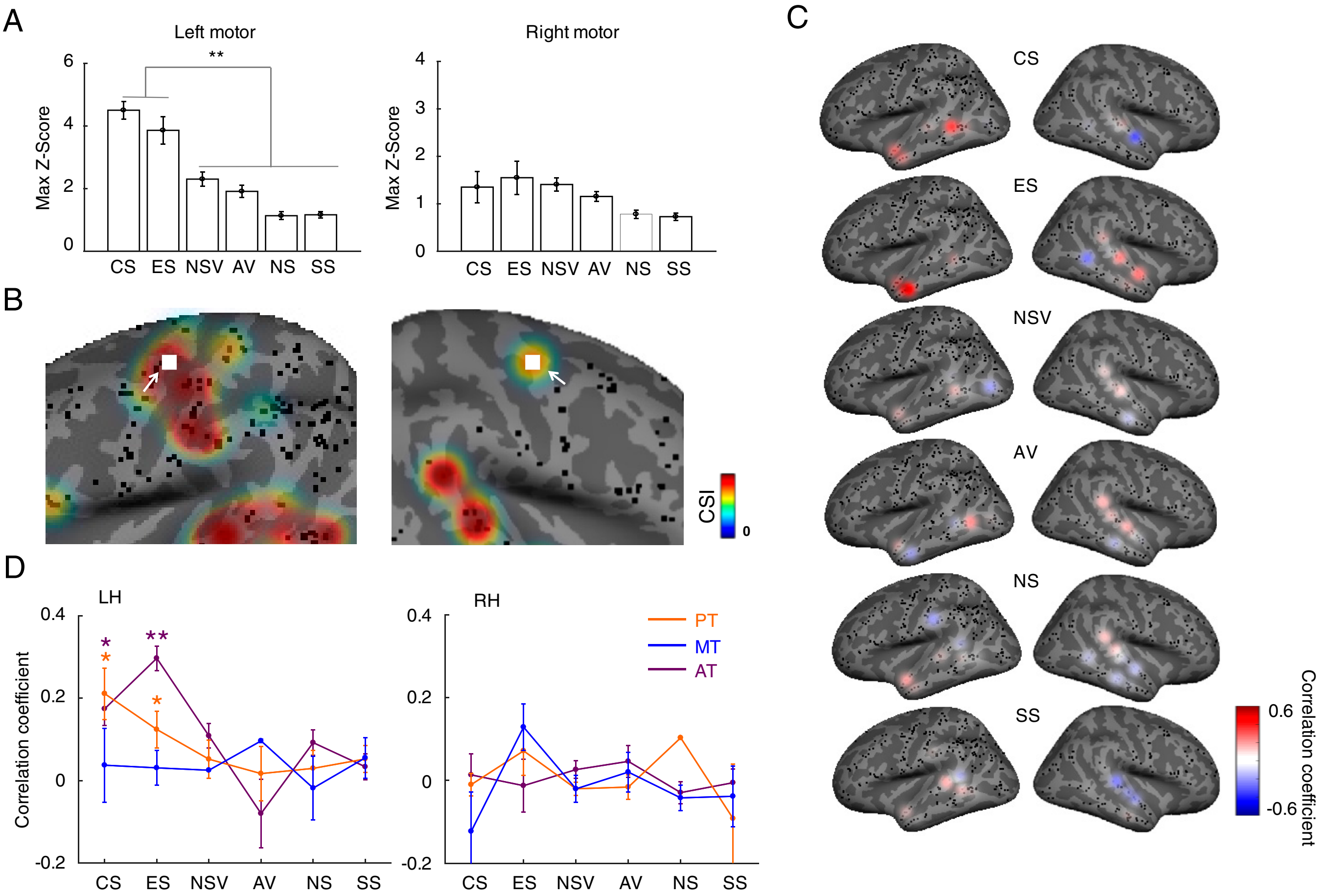

Fig. 5. Involvements of motor areas in voice processing. (*A*) Maximum Z-score responses (mean ± SEM) averaged across all electrodes of left and right motor areas in response to the six categories of sounds (**$P < 0.01$, rank sum test). (*B*) Locations of example seed electrodes in the motor areas of the left and right hemispheres. (*C*) Correlations between the example seed electrode and electrodes from temporal voice patches in both hemispheres in response to the six categories of sounds. (*D*) Mean correlation coefficients (mean ± SEM) between all seed electrodes in the motor areas and all electrodes in the temporal voice patches (*$P < 0.05$; **$P < 0.01$, permutation test).

were spoken by a single voice compared to when they were spoken by different voices, suggesting that right anterior areas have an important role in the representation of human voice identity. These results are consistent with our findings that the right AT voice patches only selectively respond to human VS (CS and ES). Similarly, in this study, left AT voice patches only show selective responses to CS, consistent with the idea that the left AT lobe is a hub for semantic processing (29). The functional roles of human AT areas are also revealed by lesion studies (30–32) and functional imaging studies (33–35). Our results on the anterior voice-specific clusters fit the notion that the AT lobe areas are involved in voice identity and semantic processing (8). Taken together, this study extends previous findings of voice stimuli processing from individual voice-specific areas of the temporal lobe to a network of voice patches with a distinct role for each patch in the processing of different voice properties. Further studies are needed to provide detailed neurophysiological evidence to clearly elucidate the functional role of each voice patch.

Dual Hierarchical Streams of Human Auditory Cortex. Similar to the visual system (36), two parallel processing streams have been proposed in the primary auditory cortex of nonhuman primates, which include a ventral "what" pathway and a dorsal "where" pathway (3). In humans, a similar dual-stream model has been proposed for speech processing, with the dorsal stream suggesting an auditory–motor integration, which is different from the dorsal "where" stream in nonhuman primates (2, 5, 6). Our study also suggests a dual-directional hierarchical information flow for processing voices in the human brain. The information flow starts from the MT voice patch and moves in two directions, one from the MT to the AT voice patch and the other one from the MT to the PT voice patch. We observed a decrease in the number of preferred voice categories (from voice to human voice to speech) and an increase in latencies along the information flow. Compared to the speech processing dual-stream model, our information flow also provides information regarding how other categories of sounds are processed. Furthermore, the proposal of the dual-stream model in humans was mainly based on functional imaging data; our proposed model is supported by neurophysiological evidence showing both the dynamics of the dual-directional information flow and the connectivity between the voice patches along the flow.

Characterizations of response properties of the voice patches along the dual-directional hierarchical information flow showed a decrease in high gamma response amplitudes and an increase in response sparseness. The changes of response amplitude and sparseness along the dual information flow may be due to increased vulnerability of higher cognitive state and that higher levels of the dual information flow are more likely to be affected by top–down signals, while lower levels can represent stimulus properties with more robust and higher high gamma

https://doi.org/10.1073/pnas.2113887118

responses. Previous studies have shown that cortical activities can be largely inhibited by top–down signals, especially attention effects (37). An alternative explanation for the changes of response amplitude and sparseness along the dual information flow is sparse coding. Sparse coding is known to be computationally efficient for higher visual processing (21), and this strategy may also be involved in higher auditory processing, which may account for the increase in sparseness along the dual information flow.

Connectivity of Voice Patches under Resting and Task-Engaged States. Resting state functional connectivity has been widely accepted and used to assess structural connectivity of different brain regions in functional imaging studies (23, 38). Slow fluctuations of resting state ECoG high gamma band activity show high correlations with fMRI resting state connectivity (22, 39). Therefore, in this study, we calculated the correlations of resting state ECoG high gamma band activity of voice patches to assess the intrinsic connectivity between them. We calculated the connectivity of within-patch electrode pairs, between-patch electrode pairs, and outside-patch electrode pairs. We found that the correlations of within-patch and between-patch electrode pairs are significantly higher than those of outside-patch electrode pairs (Fig. 4), indicating that voice patches are interconnected with each other. Furthermore, we found similar patterns of connectivity under resting and task-engaged states of all voice patches, suggesting that the functional connectivity of the voice patches are based on their intrinsic structural connectivity. This finding is consistent with the previous findings showing similar patterns of intrinsic and task-evoked brain network architectures (22, 40, 41). In this study, correlations across trials were used to infer the connectivity within the voice patch system under task-engaged states. However, this method lacks the information regarding the processing stage of each voice patch, especially in the case of parallel and/or recurrent processing. In future studies, effective connectivity analyses within this system (for example, Granger causality analyses) need to be carried out to explicitly unveil the relations within the voice patch system.

Attentional modulation of neural activity in different tasks has been found in multiple sensory modalities. In the visual system, enhanced neural activations in specific visual cortical regions have been found as a function of the attended visual attributes such as shape, color, velocity, face, and location (42–46). In the auditory system, neural activity has also been found to be modulated by attentional effects. For example, Von Kriegstein and colleagues asked subjects to selectively attend to different features of speech (identity or content) while identical stimuli were presented (47). They found increased neural activity in cortical regions that were selective to the attended speech features. In addition, such enhanced neural responses can be used to decode speech content and identity (48). In the present study, subjects were asked to perform a task to determine whether the sound they heard was a human voice or not. This particular task could potentially boost neural responses to human voice compared with other categories of sounds, which may lead to enhanced voice selectivity. We suspect that lower voice selectivity would be found if subjects perform passive listening tasks.

Functional Lateralization in Voice Processing. The anatomic and physiological asymmetries of the afferent pathways have provided the basis for the functional asymmetries of the auditory cortex (49). Two models have been proposed to explain the origin of the lateralization. One model proposes that the lateralization derives from the differences of information being processed in different hemispheres: the left hemisphere processes temporal features, whereas the right hemisphere processes spectral features (50, 51). Another model proposes that the lateralization is due to the differences of stored representations in different hemispheres: the left hemisphere stores lexical information, whereas the right hemisphere stores affective prosodic information (5).

In this study, we observed significant differences between two hemispheres in two aspects. First, significant left lateralization of native language processing was observed by comparing the selectivity to CS versus ES of all responsive electrodes in both hemispheres (*SI Appendix*, Fig. S3). These results suggest that the left AT voice patch acts as the hub for semantic processing, which is in accordance with the previous studies (8, 29). Second, we observed that the motor areas in the left hemisphere have significantly higher responses to speech sounds (CS and ES) compared to the motor areas in the right hemisphere, and significant correlations were observed between motor areas and higher order temporal voice patches (the left PT and the left AT) in response to speech sounds in the left hemisphere only (Fig. 5). Previous studies have shown that motor areas are activated during passive sound–listening tasks. Two theories (speech sensory–motor integration theory and auditory mirror neuron system theory) have been proposed to explain this finding (52–57). Our results suggest a critical role of motor areas in voice processing and the left hemisphere dominance in auditory–motor interaction in speech perception, consistent with the functional role of the dorsal stream in a previously proposed speech processing model (5, 53).

Materials and Methods

Participants. Five patients (*SI Appendix*, Fig. S1, see Table 1 for additional information, ages: 32, 45, 27, 47, and 21) with intractable epilepsy were included in this study. They were all implanted with subdural ECoG electrode grids (4-mm electrode diameter and 1-cm interelectrode center-to-center distance) as a part of clinical treatments of epilepsy. All subjects are right-handed with a normal intelligence quotient and normal hearing. They all speak Chinese as their first language and have no English background. Written informed consent was obtained from each subject before enrollment. The experiment protocol was approved by the Institutional Review Board at Tsinghua University, the affiliated Yuquan Hospital, and General Hospital of People's Liberation Army.

Stimuli and Tasks. Stimuli consisted of six categories of sounds (*SI Appendix*, Fig. S2 for details): VS (including CS and ES, both of which were single words recorded from Chinese and English native speakers in soundproofed chamber), NSV, AV, NS, and SS (phase scramble of the VS to preserve the spectral content). NSV, AV, and NS stimuli were derived from the "Animal, Artificial, Natural, Speech, and Vocal Non-Speech sounds" dataset (58). CS and ES stimuli were recorded from normal adults in a soundproof chamber using Tucker-Davis Technologies RZ6 (http://www.tdt.com). We chose Chinese words with falling–rising tones as the CS stimuli since it is easier for Chinese subjects to

Table 1. Clinical profiles of the subjects

Subject	Age	Gender	Seizure focus location	Electrode placement	Language dominance
1	32	Male	Left frontal lobe	Left FL, TL	Left
2	45	Male	Left hippocampus	Right FL, TL	Left
3	27	Male	Left hippocampus	Left PL, TL	Left
4	47	Female	Right mesial temporal lobe	Right TL, OL	Left
5	21	Male	Left occipital lobe	Left TL, PL, OL	Left

Abbreviations: TL, temporal lobe; PL, parietal lobe; OL, occipital lobe; FL, frontal lobe.

identify. ES stimuli were simple English words with consonant and vowel combinations. All stimuli had roughly similar durations (0.7 ± 0.16 s). VS had three stimuli for CS and three stimuli for ES. For the other four categories, each category had six stimuli. All stimuli were normalized to the same amplitude level. The presentation of the normalized stimuli was controlled by MATLAB (The MathWorks) using Psychophysics Toolbox 3.0 extension (59) and were delivered via inserted air-conduction earphones (ER2, Etymotic Research). The volume was adjusted to a comfortable level, ∼65 dB sound pressure level (SPL). All stimuli were presented in a randomized order with each repeated for 20 times. Subjects were asked to determine whether the sound they heard was a human voice or not by pressing a button in each trial, after the stimulus of this trial was over, using the hand ipsilateral to the side of the electrodes' coverage (sound–listening tasks).

Electrophysiological Data Acquisitions. ECoG signals were recorded via a 96-channel g.USBamp amplifier/digitizer system (g.tec) from implanted subdural electrodes with a high-pass filter of 0.01 Hz cutoff frequency, a notch filter at 50 Hz, and a sampling rate of 1,200 Hz. Four electrodes that were placed on the inner surface of the skull were used as ground and reference (two as ground and another two as reference, for redundancy). Resting state ECoG data were recorded during continuous periods of eyes opened rest.

Electrode Localizations. The locations of electrodes relative to the cortical surface were determined using Freesurfer (https://surfer.nmr.mgh.harvard.edu/). An individual three-dimensional brain with electrodes on the surface was reconstructed by aligning the presurgical high resolution T1-weighted MRI obtained by a Philips Achieva 3.0T TX scanner with the postsurgical computed tomography (CT) images obtained by the Siemens SOMATOM Sensation 64 CT. This registration was visually verified and manually adjusted when needed. In order to show all subjects' implanted electrodes on one average brain surface, we coregistered the individual MRI to the fsaverage brain by Freesurfer. All electrodes were displayed on the three-dimensional–constructed cortical surface of the average brain. Furthermore, these electrodes were also superimposed onto the inflated average brain for visualization (60–62).

ECoG Signal Preprocessing and Z-Score Calculations. All analyses were performed using MATLAB. Each channel was visually inspected for artifacts. Channels with epileptiform activity were excluded from further analysis. Notch filters from fieldtrip (https://www.fieldtriptoolbox.org/) were used to remove 50 Hz noise and its second and third harmonics. The data were down sampled at 500 Hz and then segmented into a 200-ms prestimuli baseline and an 800-ms poststimuli interval. All analyses were focused on high gamma band activities (70 to 140 Hz), which have been shown to be highly correlated with fMRI BOLD signals and population spike activities (63, 64). High gamma band activities are also the most stable responses to auditory stimuli compared to other frequency bands (65). The Z-scores of high gamma band were estimated with the following steps: 1) a 100-ms moving window (20-ms step) was used to perform short-time Fourier transform for the preprocessed ECoG signals , 2) each frequency component time series was then normalized to its own baseline mean and divided by its own baseline STD to get its own Z-score time series, and finally, 3) Z-score time series of each frequency component inside the range of 70 to 140 Hz were averaged together, producing a single Z-score time series for each trial of each channel. These procedures aimed to cancel the 1/frequency decay of power in the spectrum. A single electrode was considered as a significant responsive electrode if the maximum mean Z-scores across trials in response to either stimulus exceeded 2.

CSI. To characterize whether an electrode has sound category preference, we defined CSI as the following:

$$CSI = \frac{r^+ - r^-}{r^+ + r^-}\ ; \quad r^+ = \mathrm{mean}(r_i) \mid r_i > \bar{r}_i \ ; \ r^- = \mathrm{mean}(r_i) \mid r_i < \bar{r}_i,$$
$$i : \text{sound category index}\ ; \quad \bar{r}_i : \text{mean response of all categories}$$

where r^+ is the mean response amplitude of selective categories ($r_i > \bar{r}_i$), and r^- is the mean response amplitude of nonselective categories ($r_i < \bar{r}_i$). CSI represents the response distance between the selective categories and nonselective categories in each electrode. For visualization, the CSI values were all mapped onto the fsaverage-inflated brain using MATLAB.

CS Selectivity Index (SI$_{\text{(CS versus ES)}}$). SI$_{\text{(CS versus ES)}}$ is defined as $(R_{CS} - R_{ES})/(R_{CS} + R_{ES})$ to quantify the distance between responses to CS and ES. R_{CS} is the mean HG response across all CS stimuli, whereas R_{ES} is the mean HG response across all ES stimuli. SI$_{\text{(CS versus ES)}}$ is a value between −1 and 1. It is a positive value if the response to CS is higher than the response to ES, a negative value if the response to CS is lower than the response to ES.

Latency and Sparseness. Latency was measured within 800-ms after stimulus onset using the Z-score of each trial. As previously described, a trial is considered responsive if the Z-score of that trial exceeds the 95% CI of the prestimulus baseline mean and maintains for at least 100 ms (20). Latency was defined as the time point when Z-score first exceeds the 95% CI of the baseline mean of each responsive trial.

Sparseness was calculated based on the Z-score of each trial. For an electrode, the sparseness (21) of the activity was defined as the following:

$$\text{Sparseness} = \frac{1 - \frac{1}{n}\frac{\left(\sum r_i\right)^2}{\sum {r_i}^2}}{1 - \frac{1}{n}},$$

where n was the number of all trials (when comparing the sparseness of all electrodes, only trials of CS stimuli were involved because CS was the only stimulus that all electrodes responded to), and r_i equalled 1 when the i th trial was defined as the responsive trial. Otherwise, r_i equaled 0. The maximum sparseness was one when only one trial was the responsive trial, and the minimum sparseness was zero when all trials were the responsive trials.

Data Availability. The data and stimuli that support the finding of this study are openly (66) available at the Open Science Foundation repository (https://doi.org/10.17605/OSF.IO/SB496).

ACKNOWLEDGMENTS. This work was supported by a grant from Guoqiang Institute, Tsinghua University, and the National Key R&D Program of China (2017YFA0205904 to B.H.). We thank Y. L. Yan, X. P. Si, C. Song, R. Xu, D. Zhang, and H. Han for comments and discussion.

1. J. P. Rauschecker, B. Tian, Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11800–11806 (2000).
2. J. P. Rauschecker, S. K. Scott, Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat. Neurosci.* **12**, 718–724 (2009).
3. J. H. Kaas, T. A. Hackett, Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11793–11799 (2000).
4. B. Tian, D. Reser, A. Durham, A. Kustov, J. P. Rauschecker, Functional specialization in rhesus monkey auditory cortex. *Science* **292**, 290–293 (2001).
5. G. Hickok, D. Poeppel, The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
6. S. K. Scott, I. S. Johnsrude, The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* **26**, 100–107 (2003).
7. P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, B. Pike, Voice-selective areas in human auditory cortex. *Nature* **403**, 309–312 (2000).
8. C. Perrodin, C. Kayser, T. J. Abel, N. K. Logothetis, C. I. Petkov, Who is that? Brain networks and mechanisms for identifying individuals. *Trends Cogn. Sci.* **19**, 783–796 (2015).
9. C. R. Pernet *et al.*, The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage* **119**, 164–174 (2015).
10. P. Belin, R. J. Zatorre, P. Ahad, Human temporal-lobe response to vocal sounds. *Brain Res. Cogn. Brain Res.* **13**, 17–26 (2002).
11. S. Fecteau, J. L. Armony, Y. Joanette, P. Belin, Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage* **23**, 840–848 (2004).
12. C. F. Altmann, O. Doehrmann, J. Kaiser, Selectivity for animal vocalizations in the human auditory cortex. *Cereb. Cortex* **17**, 2601–2608 (2007).
13. C. I. Petkov *et al.*, A voice region in the monkey brain. *Nat. Neurosci.* **11**, 367–374 (2008).
14. S. Sadagopan, N. Z. Temiz-Karayol, H. U. Voss, High-field functional magnetic resonance imaging of vocalization processing in marmosets. *Sci. Rep.* **5**, 10950 (2015).
15. D. Y. Tsao, W. A. Freiwald, R. B. Tootell, M. S. Livingstone, A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).
16. D. Y. Tsao, S. Moeller, W. A. Freiwald, Comparing face patch systems in macaques and humans. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 19514–19519 (2008).
17. C. C. Hung *et al.*, Functional mapping of face-selective regions in the extrastriate visual cortex of the marmoset. *J. Neurosci.* **35**, 1160–1172 (2015).
18. S. Moeller, W. A. Freiwald, D. Y. Tsao, Patches with links: A unified system for processing faces in the macaque temporal lobe. *Science* **320**, 1355–1359 (2008).
19. P. Belin, C. Bodin, V. Aglieri, A "voice patch" system in the primate brain for processing vocal information? *Hear. Res.* **366**, 65–74 (2018).
20. K. V. Nourski *et al.*, Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *Neuroimage* **101**, 598–609 (2014).
21. W. E. Vinje, J. L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
22. B. L. Foster, V. Rangarajan, W. R. Shirer, J. Parvizi, Intrinsic and task-dependent coupling of neuronal population activity in human parietal cortex. *Neuron* **86**, 578–590 (2015).
23. M. P. van den Heuvel, R. C. Mandl, R. S. Kahn, H. E. Hulshoff Pol, Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Hum. Brain Mapp.* **30**, 3127–3141 (2009).

24. M. D. Fox, M. E. Raichle, Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8**, 700–711 (2007).
25. Y. Nir *et al.*, Interhemispheric correlations of slow spontaneous neuronal fluctuations revealed in human sensory cortex. *Nat. Neurosci.* **11**, 1100–1108 (2008).
26. Y. Yan *et al.*, Human cortical networking by probabilistic and frequency-specific coupling. *Neuroimage* **207**, 116363 (2020).
27. N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
28. P. Belin, R. J. Zatorre, Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* **14**, 2105–2109 (2003).
29. K. Patterson, P. J. Nestor, T. T. Rogers, Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* **8**, 976–987 (2007).
30. J. McGlone, Speech comprehension after unilateral injection of sodium amytal. *Brain Lang.* **22**, 150–157 (1984).
31. M. L. Gorno-Tempini *et al.*, Cognition and anatomy in three variants of primary progressive aphasia. *Ann. Neurol.* **55**, 335–346 (2004).
32. D. Poeppel, Pure word deafness and the bilateral processing of the speech code. *Cogn. Sci.* **25**, 679–693 (2001).
33. C. Humphries, T. Love, D. Swinney, G. Hickok, Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Hum. Brain Mapp.* **26**, 128–138 (2005).
34. R. Vandenberghe, A. C. Nobre, C. J. Price, The response of left temporal cortex to sentences. *J. Cogn. Neurosci.* **14**, 550–560 (2002).
35. C. Price, G. Thierry, T. Griffiths, Speech-specific auditory processing: Where is it? *Trends Cogn. Sci.* **9**, 271–276 (2005).
36. M. Mishkin, L. G. Ungerleider, K. A. Macko, Object vision and spatial vision: Two cortical pathways. *Trends Neurosci.* **6**, 414–417 (1983).
37. M. S. Beauchamp, R. W. Cox, E. A. DeYoe, Graded effects of spatial and featural attention on human area MT and associated motion processing areas. *J. Neurophysiol.* **78**, 516–520 (1997).
38. M. D. Greicius, K. Supekar, V. Menon, R. F. Dougherty, Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cereb. Cortex* **19**, 72–78 (2009).
39. C. J. Keller *et al.*, Neurophysiological investigation of spontaneous correlated and anticorrelated fluctuations of the BOLD signal. *J. Neurosci.* **33**, 6333–6342 (2013).
40. S. M. Smith *et al.*, Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13040–13045 (2009).
41. M. W. Cole, D. S. Bassett, J. D. Power, T. S. Braver, S. E. Petersen, Intrinsic and task-evoked network architectures of the human brain. *Neuron* **83**, 238–251 (2014).
42. M. Corbetta, F. M. Miezin, S. Dobmeyer, G. L. Shulman, S. E. Petersen, Attentional modulation of neural processing of shape, color, and velocity in humans. *Science* **248**, 1556–1559 (1990).
43. K. M. O'Craven, B. R. Rosen, K. K. Kwong, A. Treisman, R. L. Savoy, Voluntary attention modulates fMRI activity in human MT-MST. *Neuron* **18**, 591–598 (1997).
44. J. V. Haxby *et al.*, The functional organization of human extrastriate cortex: A PET-rCBF study of selective attention to faces and locations. *J. Neurosci.* **14**, 6336–6353 (1994).
45. E. Wojciulik, N. Kanwisher, J. Driver, Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. *J. Neurophysiol.* **79**, 1574–1578 (1998).
46. N. Kanwisher, E. Wojciulik, Visual attention: Insights from brain imaging. *Nat. Rev. Neurosci.* **1**, 91–100 (2000).
47. K. von Kriegstein, E. Eger, A. Kleinschmidt, A. L. Giraud, Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res. Cogn. Brain Res.* **17**, 48–55 (2003).
48. E. Formisano, F. De Martino, M. Bonte, R. Goebel, "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* **322**, 970–973 (2008).
49. R. J. Zatorre, P. Belin, V. B. Penhune, Structure and function of auditory cortex: Music and speech. *Trends Cogn. Sci.* **6**, 37–46 (2002).
50. R. J. Zatorre, P. Belin, Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* **11**, 946–953 (2001).
51. P. Albouy, L. Benjamin, B. Morillon, R. J. Zatorre, Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* **367**, 1043–1047 (2020).
52. A. M. Liberman, I. G. Mattingly, The motor theory of speech perception revised. *Cognition* **21**, 1–36 (1985).
53. G. B. Cogan *et al.*, Sensory-motor transformations for speech occur bilaterally. *Nature* **507**, 94–98 (2014).
54. Y. Du, B. R. Buchsbaum, C. L. Grady, C. Alain, Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7126–7131 (2014).
55. G. Rizzolatti, L. Craighero, The mirror-neuron system. *Annu. Rev. Neurosci.* **27**, 169–192 (2004).
56. E. Kohler *et al.*, Hearing sounds, understanding actions: Action representation in mirror neurons. *Science* **297**, 846–848 (2002).
57. G. Galati *et al.*, A selective representation of the meaning of actions in the auditory mirror system. *Neuroimage* **40**, 1274–1286 (2008).
58. A. Capilla, P. Belin, J. Gross, The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cereb. Cortex* **23**, 1388–1395 (2013).
59. D. H. Brainard, The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
60. Y. Zhang *et al.*, The roles of subdivisions of human insula in emotion perception and auditory processing. *Cereb. Cortex* **29**, 517–528 (2019).
61. Y. Ding *et al.*, Neural correlates of music listening and recall in the human brain. *J. Neurosci.* **39**, 8112–8123 (2019).
62. N. Guo *et al.*, Speech frequency-following response in human auditory cortex is more than a simple tracking. *Neuroimage* **226**, 117545 (2021).
63. R. Mukamel *et al.*, Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science* **309**, 951–954 (2005).
64. Y. Nir *et al.*, Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Curr. Biol.* **17**, 1275–1285 (2007).
65. N. E. Crone, D. Boatman, B. Gordon, L. Hao, Induced electrocorticographic gamma activity during auditory perception. Brazier Award-winning article, 2001. *Clin. Neurophysiol.* **112**, 565–582 (2001).
66. Y. Zhang, Hierarchical cortical networks of "voice patches" for processing voices in human brain. Open Science Framework. 10.17605/OSF.IO/SB496. Deposited 8 October 2021.