Neural correlates of sine wave speech intelligibility in human frontal and temporal cortex

Sattar Khoshkhoo[a*], Matthew K. Leonard[b,c,d*], Nima Mesgarani[e], and Edward F. Chang[b,c,d]

[a] School of Medicine, University of California, San Francisco, 505 Parnassus Ave., San Francisco, CA 94143
[b] Department of Neurological Surgery, University of California, San Francisco, 505 Parnassus Ave., San Francisco, CA 94143
[c] Center for Integrative Neuroscience, University of California, San Francisco, 675 Nelson Rising Ln., Room 535, San Francisco, CA 94158
[d] Weill Institute for Neurosciences, University of California, San Francisco, 675 Nelson Rising Ln., Room 535, San Francisco, CA 94158
[e] Department of Electrical Engineering, Columbia University, Mudd Building, Room 1339, 500 W 120th St., New York, NY 10027

[*] Equally contributing co-first authors

Corresponding Author: Edward F. Chang, Edward.Chang@ucsf.edu.

Abstract

Auditory speech comprehension is the result of neural computations that occur in a broad network that includes the temporal lobe auditory cortex and the left inferior frontal cortex. It remains unclear how representations in this network differentially contribute to speech comprehension. Here, we recorded high-density direct cortical activity during a sine wave speech (SWS) listening task to examine detailed neural speech representations when the exact same acoustic input is comprehended versus not comprehended. Listeners heard SWS sentences (pre-exposure), followed by clear versions of the same sentences, which revealed the content of the sounds (exposure), and then the same SWS sentences again (post-exposure). Across all three task phases, high-gamma neural activity in the auditory cortex superior temporal gyrus was similar, distinguishing different words based on bottom-up acoustic features. In contrast, frontal regions showed a more pronounced and sudden increase in activity only when the input was comprehended, which corresponded with stronger representational separability among spatiotemporal activity patterns evoked by different words. We observed this effect only in participants who were not able to comprehend the stimuli during the pre-exposure phase, indicating a relationship between frontal high-gamma activity and speech understanding. Together, these results demonstrate that both frontal and temporal cortical networks are involved in spoken language understanding, and that under certain listening conditions, frontal regions are involved in discriminating speech sounds.

Keywords: speech; sine-wave speech; electrocorticography; language; perception; comprehension; auditory

1. Introduction

Sometimes we hear speech without comprehending it. Although the auditory system receives input, the neural processes involved in comprehension are not engaged, and we do not access phonological, lexical, or semantic information. Yet at other times, we can listen to the same acoustic input and understand it with little effort. This distinction has been leveraged in studies of speech comprehension that use acoustically-degraded speech signals including sine-wave speech (SWS; (Remez et al., 1981)) and noise-vocoded speech (Davis and Johnsrude, 2003). Both of these stimuli manipulate the acoustic structure of speech to render it unintelligible if the listener is naïve to the stimuli. However, listeners can learn to comprehend these sounds with exposure or instruction (Dahan and Mead, 2010; Sohoglu and Davis, 2016). This paradigm allows for the direct comparison of the same acoustic stimulus under two conditions of intelligibility and unintelligibility.

Previous studies have discovered robust activation of the left inferior frontal and pre-motor regions when acoustically degraded speech becomes intelligible (Davis and Johnsrude, 2003; Eisner et al., 2010; Hervais-Adelman et al., 2012). Other studies have found changes in activity related to comprehension in secondary auditory regions including the superior temporal gyrus (STG; (Giraud et al., 2004; Holdgraf et al., 2016; Sohoglu et al., 2012)), possibly related to top-down modulation by linguistic knowledge. In particular, the relative engagement of frontal and temporal language regions during degraded speech understanding may be influenced by a number of factors including comprehension ability (Eisner et al., 2010; Möttönen et al., 2006), working memory (Eisner et al., 2010), acoustic complexity (Sohoglu and Davis, 2016), and neural factors such as lateralization (Giraud et al., 2004; Sohoglu and Davis, 2016).

To date, it is not clear what actual information is being processed in these temporal and frontal cortical regions at a representational level. To address this question, we examined the fine-scale neurophysiological changes associated with comprehension of acoustically-degraded sine-wave speech. We used intracranial direct neural recordings in three human participants to examine cortical activity while participants listened to sentences during three task phases: (1) Pre-exposure, where the SWS stimulus is not comprehended, (2) Exposure, where participants listen to the original clear speech (CS) versions of the stimuli, and (3) Post-exposure, where the same SWS stimulus is immediately intelligible after the exposure phase.

The spatiotemporal resolution of electrocorticography (ECoG) activity allowed us to identify the location and rapid time-course processing during comprehension-related neural activity. Equally important, using clustering analyses on population neural data, we determined that both auditory and frontal regions represent information that discriminates different words under conditions where the bottom-up speech input is comprehended.

2. Materials and Methods

2.1 Participants

Three human participants undergoing clinical work-up for epilepsy surgery were implanted with high-density (4mm pitch) multi-electrode (256 channels) cortical surface arrays. Arrays were placed over the lateral surface of the language dominant hemisphere (confirmed by the Wada test), covering superior temporal gyrus, middle temporal gyrus, pre-central gyrus, post-central gyrus, and posterior inferior frontal gyrus (**Fig. S1**). Two participants were left hemisphere language dominant (one post-CS comprehender and one pre-CS comprehender) and one participant was right hemisphere language dominant (in the post-CS comprehender group). Analyses of data from other experiments and comparisons with left-dominant participants did not indicate any major or systematic differences in cortical organization due to right hemisphere dominance in in this participant. Information about each participant is shown in **Table 1**. Note that although participant 1 showed possible epileptiform activity in posterior temporal cortex, the data were carefully cleaned to remove any possible spiking or other seizure-related artifacts, and we have not observed differences in this participant's data across other tasks completed during ECoG monitoring. Nonetheless, it is important to note that data from participants with epilepsy may be qualitatively different compared to individuals who do not have epilepsy, and should therefore be considered in the context of research with healthy participants (see Discussion). Finally, although the present study was conducted with only three participants, we examined all effects within-subject, as is common practice in neurophysiological experiments with humans and non-human primates, where larger samples are often unattainable.

Table 1: Participant characteristics.

Participant	Age	Sex	Handedness	Seizure Focus
1	44	M	Right	Unclear; possible activity in posterior temporal cortex
2	29	F	Left	Right mesial temporal lobe

3	25	F	Right	Left mesial temporal lobe

2.2 Stimuli and Task design

Participants listened to spoken sentences from a public database called Coordinate Response Measure (CRM, Bolia et al., 2000), commonly used in communication research. Each sentence consisted of a call sign, a color, and a number, all embedded in a carrier phrase, for example “ready *ringo* go to *blue two* now.” Two call signs (ringo or tiger) were paired with six color-number combinations (blue 2, blue 5, red 2, red 7, green 5, green 7) for 12 unique phrases. Each set of phrases was spoken by a male and a female speaker (speakers one and five in the CRM corpus) resulting in 24 sentences.

These sentences were presented in two different forms during three task phases. Clear speech (CS) samples were the unaltered forms of the sentences. Sine wave speech (SWS) samples were created from the CS samples using a publicly available MATLAB (MathWorks, Inc.) program (SineWave Synthesis package, Philip Rubin, Steve Frost, and Dan Ellis, Haskins Laboratories, New Haven, CT, USA) to generate 24 SWS samples (Liebenthal et al., 2001; Remez et al., 1981). This program uses linear predictive coding analysis to find formant center frequencies and amplitudes and returns up to 4 sinusoids based on those parameters. Example acoustic waveforms and auditory spectrograms for SWS and CS are shown in **Fig. 1Ai** and **Fig. 1Bi**, respectively. The resulting SWS sounds have limited broadband formants or short-timescale cues that are characteristic of many discriminative features of speech sounds, making them unlike natural speech in terms of their acoustic properties.

During ECoG recording, participants listened to 24 CS or SWS speech samples, each repeated 3-5 times during a given experimental block. Speech samples were shuffled such that each variable (call sign, color, number) was presented the same number of times. The sounds were presented diotically from a loudspeaker placed ~1m in front of the participant.

The task consisted of three phases: pre-exposure (96-144 trials), exposure (67-101 trials), and post-exposure (73-164 trials). In the pre-exposure phase, participants listened to SWS samples without any instructions about the source or the content of the sounds. To ensure that they were paying attention, participants were asked to describe what they heard in each phase. Two of the participants did not understand SWS during pre-exposure (post-SWS comprehenders) and described it as “sounds more like birds than dolphins” and “there was a little bit of words…it sounded like the same thing over and over.” The third participant quickly deciphered the SWS samples and called it “electronic voice…reading out chess moves…speech-like, but clear words.” In the exposure phase, participants listened to CS versions of the same sentences, and on each trial reported the color and number using a customized graphical user interface in MATLAB. After being exposed to the CS stimuli, participants were instructed that the SWS samples were the acoustically modified versions of the CS that they were listening to, and then they completed the rest of the block. All participants reported the color and number combinations of the CS samples with 100% accuracy (**Fig. S2**). Finally, in the post-exposure phase, they listened to the SWS samples again while reporting the color/number combinations using the GUI. The post-exposure phase consisted of several experimental blocks—variable across participants—until the participant responded with >95% accuracy. The post-exposure data used for analyses are from the last experimental block (i.e., when SWS is understood accurately and with high confidence).

2.3 Data acquisition and preprocessing

Electrocorticographic (ECoG) signals were recorded at 3,052 Hz using a multichannel amplifier optically connected to a digital signal acquisition system (Tucker-Davis Technologies, Alachua, FL, USA). Subsequently, broadband ECoG signals were visualized and electrodes with excessive artifacts or noise were excluded from the final analysis. Notch filters were applied to remove line noise (60, 120, and 180 Hz) and then non-neural noise from the amplifier was removed using common-average referencing (CAR). CAR was taken across 16 channel blocks and significantly reduced artifacts while preserving the neural signal. High-gamma band signals were extracted using the Hilbert transform by averaging the analytic amplitude of 8 semi-logarithmically spaced bands from 70 to 150 Hz (Bouchard et al., 2013; Crone et al., 1998a), then downsampled to 100Hz and z-scored relative to the pre-trial baseline for each electrode. Finally, the high-gamma data were segmented into epochs spanning 1500ms before onset to 1000ms after offset of each sentence.

2.4 Electrode selection

Electrode selection was based on two criteria: physical location on the brain surface and overall responsiveness to SWS. Electrodes over pre-central and inferior frontal gyri were labelled frontal electrodes (n = 115 electrodes for post-CS comprehenders, and n = 50 electrodes for the pre-CS comprehender) and infra-Sylvian (temporal lobe) electrodes primarily overlying the superior temporal gyrus were called STG electrodes (n = 85 electrodes for post-CS comprehenders, and n = 35 electrodes for the pre-CS comprehender).

For multidimensional scaling (MDS) analyses shown in Fig. 3 and 4, electrodes were chosen based on their responsiveness to SWS. A Student’s paired t-test was performed comparing neural responses for each electrode while listening to the SWS samples vs. the silent period preceding them. A threshold of 20 was set for the t-values to choose the most SWS-responsive electrodes. Only frontal (n = 32 electrodes for post-CS comprehenders and n = 13 for the pre-CS comprehender) and STG (n = 45 electrodes for post-CS comprehenders and n = 34 for the pre-CS comprehender) electrodes with t-values greater than 20 were included in MDS analysis. This threshold was chosen because it yielded the best separability in MDS space, however, our conclusions were not dependent upon its exact value.

2.5 Evoked response amplitude and correlation analysis

To compare evoked response amplitudes in the high-gamma range between different experimental phases (pre-exposure, exposure, and post-exposure), mean neural activity at each electrode during the silent period was subtracted from the post-stimulus response to adjust for variability in baseline. Then mean neural responses to individual stimuli for each electrode were calculated for the selected frontal and STG electrodes (Material and methods: Electrode selection), ex. **Fig. 2A** and **Fig. 2B**.

High-gamma evoked response correlation analysis was used to assess the degree of similarity between mean responses to SWS and CS. A built-in MATLAB function (corrcoeff) was used to calculate the linear correlation between mean neural responses for each of the SWS (pre- or post-exposure) vs. CS (exposure) samples. The higher the correlation coefficient, the greater the degree of similarity between the mean evoked responses.

2.6 Separability analysis

To investigate the speech features of SWS that are encoded in the brain pre- and post-exposure, we examined the relational organization of the neural responses to different words. We applied unsupervised multidimensional scaling (MDS) to the distance matrix of the mean neural responses in frontal and STG regions. All speech samples were aligned according to the onset time of the word of interest (call sign, color, number). Then, a 24x24 distance matrix was created by calculating the linear correlation between pairs of mean neural responses and then subtracting the resulting coefficient from 1. This analysis was performed on neural data in 150ms segments since that roughly equaled the length of the shortest word. MDS was then used to project these distances into a lower dimensional space for visualization. The encoding of specific speech features was examined by determining whether multiple instances of the same word were located closer to one another in the MDS space relative to the other words (**Fig. 3A** and **Fig. 4A**).

To quantify this relational organization in MDS space, a separability value (F-statistic) was calculated by dividing between-word variability over within-word variability. Since we were interested in the change in separability over the course of hearing the stimulus, we calculated separability at different time-points relative to the onset of the word of interest, which we refer to as *dynamic separability*. Dynamic separability curves were created by calculating separability values using a moving 150ms window in 10-millisecond intervals from 400ms before to 400ms after the onset of the words of interest (**Fig. 3B** and **Fig. 4B**). Pre-onset and post-onset separability were quantified using a 50ms window around the peak separability on the dynamic separability curve, before and after the onset of the words of interest (**Fig. 3C**, **Fig. 3D**, **Fig. 4C**, and **Fig. 4D**).

2.7 Statistical analysis

All statistical analyses were performed using custom-written and built-in MATLAB functions. We used one-way ANOVA with Tukey-Kramer multiple comparisons test, unless otherwise specified.

3. Results

We present the results of the experimental manipulation in two main sections. In the first section (3.1; Figures 1 and 2), we examined whether there are changes in neural activity response amplitudes across task phases, and whether these changes are different between superior temporal and frontal regions. In the second section (3.2; Figures 3 and 4), we asked whether these changes reflect differences in stimulus encoding, and whether these differences are related to comprehension. We observed a dissociation between changes in response amplitude and changes in stimulus encoding, the implications of which we will examine in the Discussion (section 4).

3.1.1 Frontal cortex activity with comprehension

Participants first listened to SWS samples without any prior information about what they were hearing (pre-exposure; **Fig. 1Ai**). Two participants reported that they did not hear anything familiar or intelligible in this condition, and neural activity in response to SWS was largely confined to STG (**Fig. 1Aii**; $p < 0.01$, Bonferroni corrected for 256 electrodes to identify only the most speech-responsive neural populations). In the representative subject shown in Figure 1, there were 29 STG and 11 supra-Sylvian (frontal/parietal) electrodes that showed significant responses. For each participant across all supra-Sylvian electrodes, we also examined whether mean high-gamma activity changed monotonically on a trial-by-trial basis throughout the pre-exposure phase, and we did not observe any such effect.

In the next phase of the task (exposure), participants listened to CS samples and were given instructions about the task (**Fig. 1Bi**). Compared to the pre-exposure phase, a similar number of STG electrodes showed significant responses (31 electrodes). However, we observed a striking increase in the number of active supra-Sylvian electrodes (36 electrodes; **Fig. 1Bii**). The crucial experimental comparison was between pre-exposure SWS and post-exposure SWS, when listeners reported understanding the stimuli up to at least 95% accuracy (**Fig. 1Ci**). We found that whereas roughly the same number of STG electrodes showed significant responses in the pre-exposure (29 electrodes) and post-exposure (30 electrodes) phases, the number of significant electrodes in supra-Sylvian cortex increased from the pre-exposure (11 electrodes) to the post-exposure (27 electrodes) phase (**Fig. 1Cii**).

Together, these results demonstrate a striking and sudden change in activity in lateral frontal regions. Critically, the number of speech-responsive electrodes in supra-Sylvian regions changed dramatically when listeners had sufficient knowledge to understand the spoken sentences, even though the acoustic input was identical.

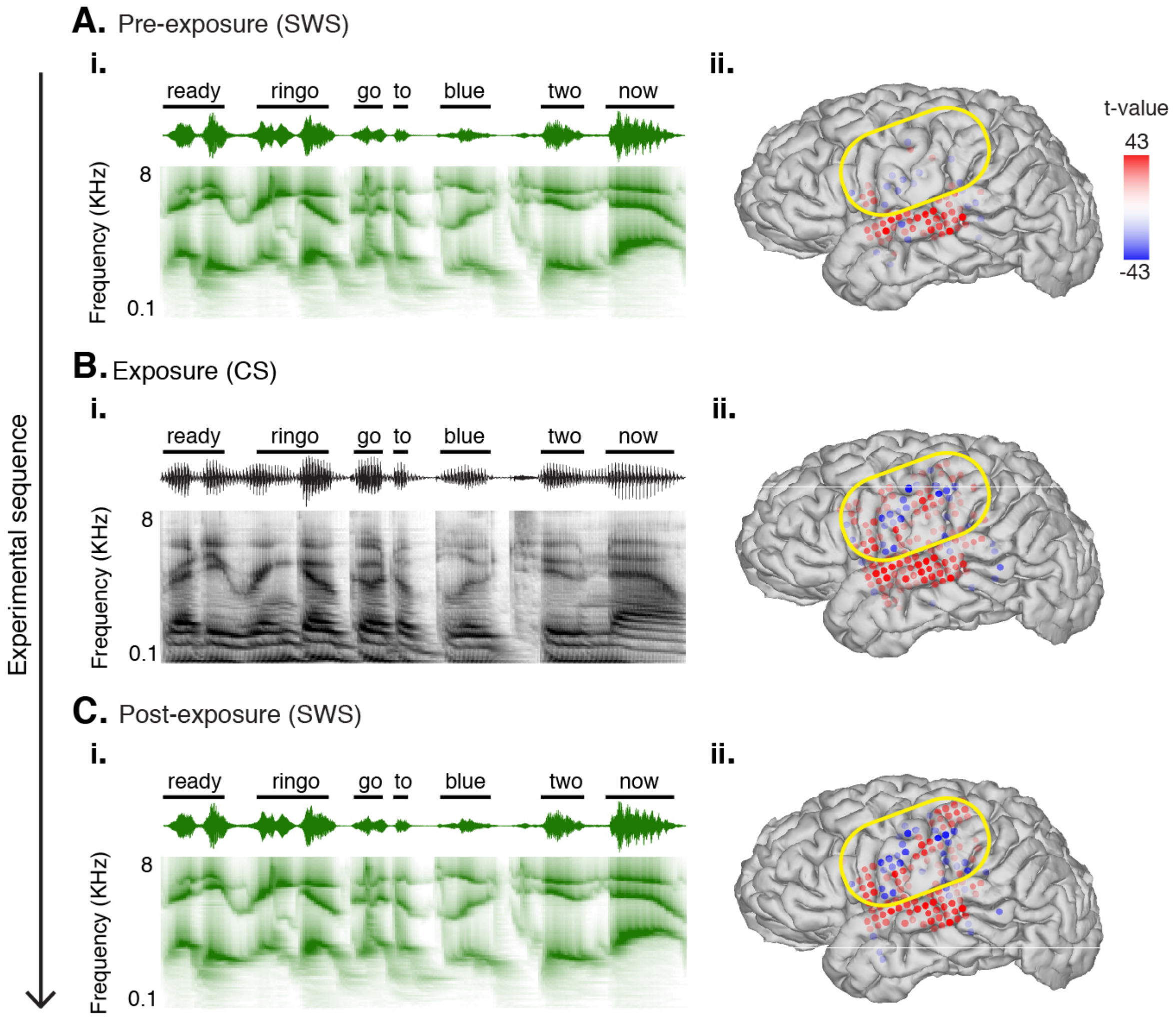

Figure 1. Frontal and parietal electrodes are activated for clear speech and post-exposure SWS. (A) (i) Example acoustic waveform and auditory spectrogram for sine-wave speech (SWS). **(ii)** Electrodes that are significantly responsive to SWS, comparing silent periods to SWS sounds in the pre-exposure period (two-tailed t-test, p<0.01 corrected for 256 electrodes). Color bar represents t-values. **(B) (i)** Example acoustic waveform and auditory spectrogram for clear speech (CS). **(ii)** Same as A(ii), but during exposure to CS. **(C)** Same as A, but in the post-exposure phase. Increased t-values in supra-Sylvian electrodes (highlighted by the yellow ovals) are apparent compared to pre-exposure SWS phase.

3.1.2 Frontal high-gamma ERP amplitude and correlation increase in the post-exposure period

Across all three participants, the increase in activity from pre-stimulus baseline was observed primarily in supra-Sylvian areas including the pre-central and inferior frontal gyri (frontal regions). Previous research has implicated these regions in speech processing (Cheung et al., 2016; Cogan et al., 2014; Edwards et al., 2010; Leonard et al., 2016; Pulvermüller et al., 2006; Wilson et al., 2004), and specifically in the comprehension of spoken input under nonideal listening conditions (Sohoglu and Davis, 2016; Sohoglu et al., 2012). We examined the specific role of these regions in the perceptual switch that occurs after the exposure phase, and characterized the nature of the speech representations in frontal and auditory cortex.

Across the three phases of the experiment, there was a clear change in the high-gamma evoked response to speech throughout the timecourse of the sentences in frontal electrodes (**Fig. 2A**). For SWS, high-gamma activity significantly increased post-exposure ($p<10^{-7}$ between pre- and post-exposure; Tukey-Kramer multiple comparisons test; n = 165 electrodes from 3 participants; **Fig. 2Bi**; see **Figure S3** for effect sizes). On average across the timecourse of speech, this higher level of activity to post-exposure SWS exceeded the response to CS in those electrodes (p = 0.006 between exposure and post-exposure; Tukey-Kramer multiple comparisons test; n = 165 electrodes from 3 participants; **Fig. 2Bi**).

Some listeners are able to understand SWS without explicit training or instruction. Two participants did not understand SWS during the pre-exposure block, but the third participant did. We therefore divided our cohort into two groups (post-CS comprehender and pre-CS comprehender), and hypothesized that the change in frontal activity would not be observed in the participant who did not exhibit a perceptual switch.

The post-CS comprehenders showed the same increase in frontal activity between pre-exposure and post-exposure as we observed at the group level ($p<10^{-8}$ between pre- and post-exposure and p<0.05 between exposure and post-exposure; Tukey-Kramer multiple comparisons test; n = 115 electrodes from 2 participants; **Fig. 2Bii**). In contrast, the pre-CS comprehender did not show any significant change in mean activity for frontal electrodes over the three phases of the experiment (p = 0.911 between pre-exposure and exposure, p = 0.515 between pre- and post-exposure, p = 0.771 between exposure and post-exposure; Tukey-Kramer multiple comparisons test; n = 50 electrodes from 1 participant; **Fig. 2Biii**). These results indicate that changes in frontal activity may be linked to whether listeners can extract meaning from the acoustic signal.

To establish that these changes in mean activity reflect the encoding of spoken content, we calculated the linear correlation between mean SWS high-gamma ERPs (pre- and post-exposure) and mean CS ERPs (exposure). This analysis specifically examines whether modulations in the speech envelope that are critical for comprehension are encoded in cortical areas implicated in understanding degraded speech. For post-CS comprehenders, the correlation between CS and post-exposure SWS was significantly higher than the correlation between CS and pre-exposure SWS ($p<10^{-6}$ via ANOVA; n = 115 electrodes from 2 participants; **Fig. 2Ci**). In contrast, the magnitude of the effect in the pre-CS comprehender was smaller, and did not reach statistical significance (p = 0.064 via ANOVA; n = 50 electrodes from 1 participant; **Fig. 2Cii**).

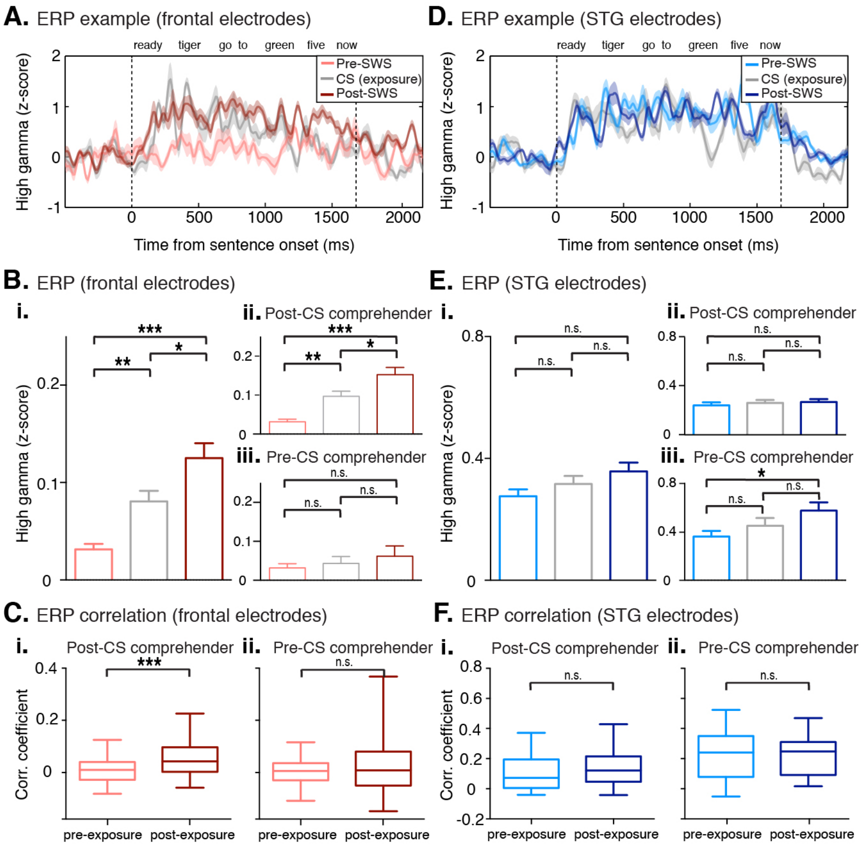

Figure 2. Frontal electrodes demonstrate increase in post-exposure high-gamma ERP amplitude and correlation (A) Example mean high-gamma event-related potentials (ERPs) for electrodes in the frontal regions for a post-CS comprehender. Dashed vertical lines mark the times of onset and offset of the auditory sample. Frontal electrodes on average show increased responses to task phases where stimuli are comprehended. **(B) (i)** Mean ERPs in frontal electrodes for all participants in response to SWS (pre- and post- exposure) and CS (exposure). **(ii)** Same as C(i), but for the post-CS comprehender group only. **(iii)** Same as C(i), but for the pre-CS comprehender only. These results show that activity increases over the different task phases are due to comprehension. **(C) (i)** Mean linear correlation coefficient between SWS (pre- and post- exposure) ERPs vs. CS (exposure) ERPs, for post-CS comprehenders only. **(ii)** Same as D(i), but for pre-CS comprehender only. **(D)** Same as A, but in the STG electrodes. STG electrodes do not show differences in overall activity across task phases. **(E)** Same as B, but in STG electrodes. **(F)** Same as C, but in STG electrodes. All data show means ± SEM and are analyzed using one-way ANOVA and multiple comparisons test. Shading in panels A and D are ± SEM. Boxes in panels C-F are ± SEM and whiskers represent 5-95 percentile over electrodes. Not significant (n.s.), p<0.05 (*), p<0.01 (**), p<0.001 (***).

3.1.3 Auditory cortex representations do not change with intelligibility of SWS

Previous work has suggested that frontal regions provide a top-down modulatory signal to auditory cortex in the temporal lobe. To evaluate this hypothesis, we examined changes in activity over the course of the experiment in STG electrodes. Strikingly, evoked responses throughout the timecourse of the sentences were highly overlapping between all three experimental phases (**Fig. 2D, Fig. 2Ei**; pre-SWS vs. CS p = 0.515; pre-SWS vs. post-SWS p = 0.067; CS vs. post-SWS p = 0.500). When we examined STG activity separately for post-CS comprehenders (**Fig. 2Eii**) and the pre-CS comprehender (**Fig. 2Eiii**), we observed only a small overall difference for the pre-CS comprehender between pre- and post-exposure SWS (p = 0.037; Tukey-Kramer multiple comparisons test; n=35 electrodes from 1 participant). There was also no difference in correlations between pre-exposure and CS, and post-exposure and CS for either the post-CS comprehenders (p = 0.111 via ANOVA; n = 85 electrodes from 2 participants; **Fig. 2Fi**) or the pre-CS comprehender (p = 0.831 via ANOVA; n = 35 electrodes from 1 participant; **Fig. 2Fii**).

Together, these results suggest that changes in the perception and comprehension of degraded speech are accompanied by concomitant changes in neural activity in pre-central and inferior frontal cortex. However, we observed that the nature of these effects, including the magnitude of the change in frontal versus STG electrodes, was somewhat sensitive to analytical factors like electrode selection. For example, by including only electrodes with significantly above-baseline responses to SWS, there was a change in STG high-gamma activity between pre-SWS and post-SWS conditions (p = 0.001, Tukey-Kramer multiple comparisons test). Although the effect was smaller than what we observed in frontal electrodes, these results indicate that simply comparing changes in the average amplitude of high-gamma activity between task phases does not explain why such changes occur, what they reflect in terms of stimulus encoding, or how they might contribute to speech comprehension. Therefore, we used multivariate analysis methods to examine the representations that high-gamma activity in each region encodes, and how these representations change over the course of the task.

3.2.1 Frontal cortex representations during comprehension

By only evaluating changes in the level of high-gamma activity at individual electrodes across the task phases, it is not possible to understand why the activity changes, or what information the activity reflects. To directly evaluate the neural processes associated with changes in comprehension, we asked whether frontal and auditory regions differentially encode word-specific information in each phase of the task. We used multi-dimensional scaling (MDS) to calculate distances between neural representations of unique words in the task (i.e., the three sets of colors, call signs, and numbers). For example, during the pre-exposure phase in a post-CS comprehender, frontal electrodes do not encode a clear separation between the three colors (red, green, and blue). However, such separation emerges once comprehension is possible during the post-exposure phase (**Fig. 3A**).

We calculated stimulus separability for frontal electrodes as a function of the time of onset for the word of interest using a moving window at 10ms intervals (‘dynamic separability curves’; Materials and Methods: Separability analysis). When spoken input is understood (e.g., CS and post-exposure SWS for post-CS comprehenders), activity in frontal regions clearly discriminates specific words (high normalized separability values after word onset in **Fig. 3Bi**). This was confirmed in the pre-CS comprehender, who showed similarly high separability for all three phases of the task (**Fig. 3Bii**).

We quantified these effects across all participants, and found that the post-CS comprehenders showed no significant increase in separability between pre-word baseline and the response to the word during the pre-exposure phase (p = 0.721 via ANOVA; n = 30 for 2 participants and 3 variables; **Fig. 3Ci;** see **Figure S4** for effect sizes), but did show significant increases during the exposure phase ($p<10^{-7}$; **Fig. 3Cii**) and post-exposure phase ($p<10^{-5}$; **Fig. 3Ciii**). Crucially, the difference in separability from pre- to post- exposure was significant for the post-CS comprehenders ($p<10^{-5}$ via ANOVA; n = 30 for 2 participants and 3 variables; **Fig. 3Civ**). Confirming that these effects reflect changes in understanding, the pre-CS comprehender showed significant increases in separability between pre-word baseline and the response to the word during the pre-exposure phase ($p<10^{-9}$ via ANOVA; n = 15 for 1 participant and 3 variables; **Fig. 3Di**), the exposure phase ($p<10^{-4}$ via ANOVA; n = 15 for 1 participant and 3 variables; **Fig. 3Dii**), and the post-exposure phase ($p<10^{-7}$ via ANOVA; n = 15 for 1 participant and 3 variables; **Fig. 3Diii**). The difference in separability from pre- to post- exposure trended toward significance for the pre-CS comprehender (p = 0.072 via ANOVA; n = 15 for 1 participant and 3 variables; **Fig. 3Div**).

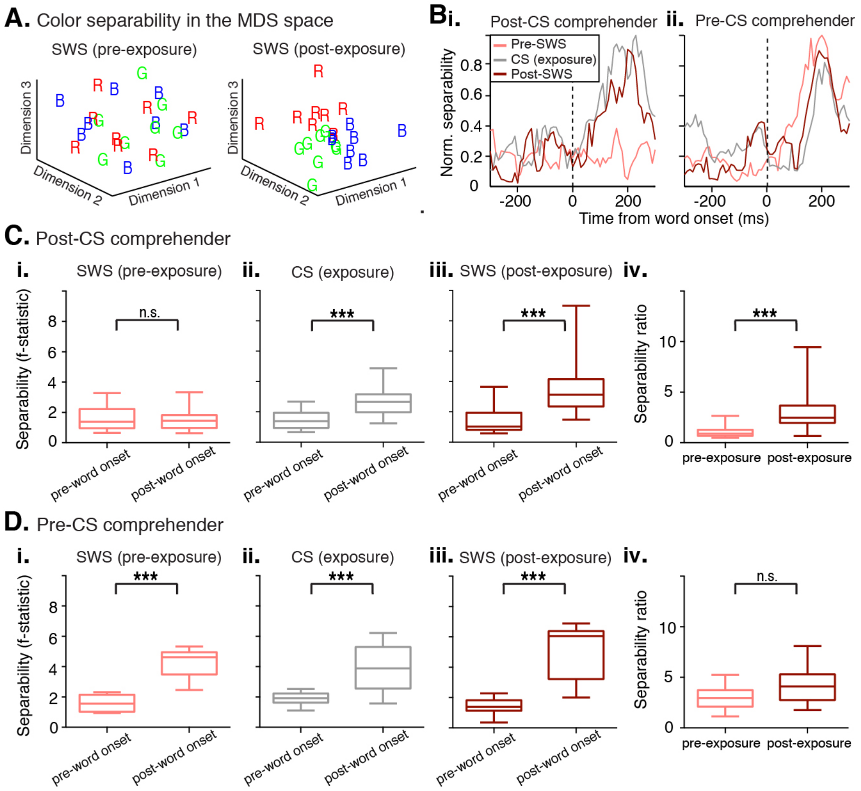

Figure 3. Exposure to CS improves SWS word separability using the frontal electrodes. (A) Example MDS representation of 'color' words from frontal electrodes during pre-exposure (left) and post-exposure (right) SWS task phases. **(B) (i)** Example dynamic separability curves relative to the onset of the word of interest (e.g., color) using the frontal electrodes for a post-CS comprehender. Each data point is the F-statistic normalized to the peak. **(ii)** Same as B(i), but for number separability in a pre-CS comprehender. These results demonstrate rapid high-gamma evoked activity changes according to comprehension. **(C) (i)** Mean separability before and after the onset of all the words of interest (call sign, color, number) using the frontal electrodes, pre-exposure, in post-CS comprehenders. **(ii)** Same as C(i), but during exposure to CS. **(iii)** Same as C(i), but post-exposure. **(iv)** Mean ratio of post-exposure to pre-exposure separability in post-CS comprehenders. **(D)** Same as C, but for the pre-CS comprehender. All data in panels C and D

show means ± SEM and are analyzed using one-way ANOVA. Boxes in panels C and D are ± SEM and whiskers represent 5-95 percentile. Not significant (n.s.), p<0.001 (***).

3.2.2 STG representations during comprehension

Finally, we calculated dynamic separability curves for STG electrodes as a function of the time of onset for the word of interest. An example of color separability in MDS space using STG electrodes is shown in **Fig. 4A**. Regardless of whether spoken input is understood (i.e., pre- vs. post- exposure), STG activity clearly separates the acoustic input for the three color words. The timecourse of separability in STG further demonstrates that high separability is apparent post-word onset for both post-CS comprehenders (**Fig. 4Bi**) and pre-CS comprehenders (**Fig. 4Bii**). This likely reflects the ability of this high-order auditory area to discriminate fine-scale acoustic differences between stimuli. Notably, the separability for SWS was significantly higher compared to CS in both participant groups (all p<0.035), which is likely due to bottom-up acoustic differences in the stimuli.

We again quantified these effects across all participants. In both post-CS comprehenders and pre-CS comprehenders, separability significantly increased from pre-word baseline to word response ($p<10^{-12}$ for post-CS comprehenders and $p<10^{-3}$ for pre-CS comprehenders via ANOVA; n = 30 for 2 participants and 3 variables for post-CS comprehenders and n = 15 for 1 participant and 3 variables for the pre-CS comprehender; **Fig. 4Ci, 4Di;** see **Figure S5** for effect sizes). Similarly, separability significantly increased from pre-word baseline to word response for both the exposure phase ($p<10^{-7}$ for post-CS comprehender and p = 0.008 for pre-CS comprehender via ANOVA; n = 30 for 2 participants and 3 variables for post-CS comprehenders and n = 15 for 1 participant and 3 variables for the pre-CS comprehender; **Fig. 4Cii, 4Dii**) and the post-exposure phase ($p<10^{-4}$ for post-CS comprehender and $p<10^{-4}$ for the pre-CS comprehender via ANOVA; n = 30 for 2 participants and 3 variables for post-CS comprehenders and n = 15 for 1 participant and 3 variables for the pre-CS comprehender; **Fig. 4Ciii, 4Diii**). In both groups, there was no change from pre-exposure to post-exposure (post-CS comprehenders: p = 0.280 via ANOVA; n = 30 for 2 participants and 3 variables; **Fig. 3Civ**; pre-CS comprehenders: p = 0.196 via ANOVA; n = 15 for 1 participant and 3 variables; **Fig. 3Div**), indicating that STG electrodes distinguish individual words equally well regardless of speech comprehension.

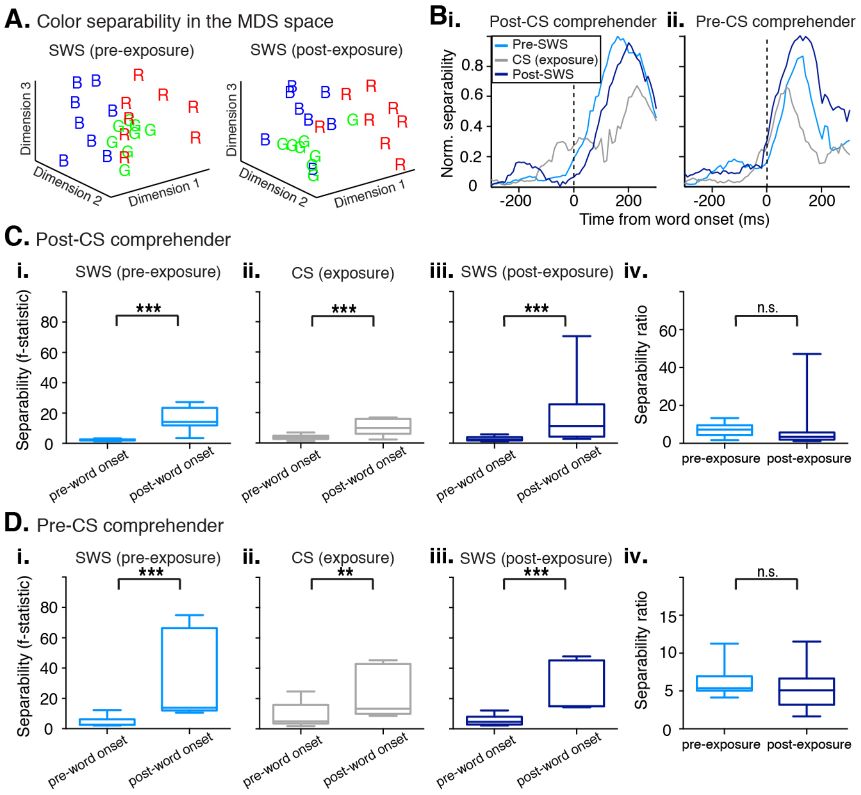

Figure 4. Exposure to CS has no effect on SWS word separability using the STG electrodes. (A) Example MDS representation of 'color' words from STG electrodes during pre-exposure (left) and post-exposure (right) SWS task phases. **(B) (i)** Example dynamic separability curves relative to the onset of the word of interest (e.g., color) using the STG electrodes for a post-CS comprehender. Each data point is f-statistic normalized to the peak. **(ii)** Same as B(i), but in a pre-CS comprehender. These results demonstrate that evoked speech representations in STG are largely unchanged across task phases, regardless of stimulus comprehension. **(C) (i)** Mean separability before and after the onset of all the words of interest (call sign, color, number) using the STG electrodes, pre-exposure, in post-CS comprehenders. **(ii)** Same as C(i), but during exposure to CS. **(iii)** Same as C(i), but post-exposure. **(iv)** Mean ratio of post-exposure to pre-exposure separability in post-CS comprehenders. **(D)** Same as C, but for the pre-CS comprehender. All data in panels C and D show means ± SEM and are analyzed using one-way ANOVA. Boxes in panels C and D are ± SEM and whiskers represent 5-95 percentile. Not significant (n.s.), p<0.01 (**), p<0.001 (***).

4. Discussion

This study investigated the neural computations and representations that are directly involved in comprehending speech. These results demonstrate that speech comprehension involves activity in multiple peri-Sylvian brain regions including superior temporal auditory and inferior frontal cortex. Most crucially, we show that frontal regions encode speech representations only when the stimuli are comprehended as spoken input. In contrast, superior temporal auditory cortical responses were very similar irrespective of intelligibility. As discussed below, these results provide important additional information about meaningful linguistic representations in these regions, and suggest that specific task and stimulus conditions may influence the degree to which neural activity in both auditory and frontal regions are modulated by higher-order representations.

Our findings are consistent with previous work that has implicated frontal cortex in the sudden perceptual shift that occurs when listeners are informed of the nature of SWS or noise-vocoded speech (Davis and Johnsrude, 2003; Eisner et al., 2010; Hervais-Adelman et al., 2012). The nature of the contributions these regions make to speech comprehension has remained unclear. It has been suggested that left inferior frontal cortex in particular is recruited under difficult listening conditions to support and augment auditory processing that occurs in superior temporal regions (Hervais-Adelman et al., 2012). Our results are consistent with recent studies showing that somato-motor regions encode higher-order abstract features of syllables, regardless of their exact acoustic properties (Evans and Davis, 2015).

Here, we demonstrate that increased activity during the post-exposure phase (indexed by high-gamma) is specifically associated with computations that discriminate task-relevant speech sounds. This specificity is possible due to the temporal and spatial resolution of ECoG, and the relatively high signal-to-noise properties of neural activity recorded directly from the cortical surface (Crone et al., 1998b; Leonard and Chang, 2016; Mesgarani et al., 2014). This resolution has allowed us to examine the nature of the ubiquitous increases in activity or activation levels that have been observed in frontal regions across most studies.

We have also demonstrated that this discriminative speech activity in frontal regions is associated with comprehension. Similar to previous studies (Möttönen et al., 2006), one participant did not experience the perceptual pop-out effect. In the present study, one participant was able to comprehend SWS during the pre-exposure phase. In this case, high-gamma activity and stimulus discriminability in frontal regions were equally high during pre- and post- exposure periods. This contrasted with post-CS comprehenders, where frontal regions showed significant increases in high-gamma activity and stimulus discriminability across the different task phases. This result also confirms that these changes do not simply reflect order effects in the structure of the task, or differences related to task-based attention. It is, however, possible that simply being able to comprehend the stimuli causes changes in participant motivation and attention, which in turn lead to enhanced encoding. The present results are unfortunately unable to address this question directly.

Previous work has identified both frontal and superior temporal regions as being important for speech comprehension in these types of experiments. Several studies have shown that pre- versus post- exposure differences are primarily localized to the superior temporal cortex (Liebenthal et al., 2005), and in particular the posterior STS (Dehaene-Lambertz et al., 2005; Möttönen et al., 2006). In some cases, these studies have used region-of-interest (ROI) analyses to focus on superior temporal regions, and therefore may not have observed changes in frontal regions related to comprehension. However, the present results examined both regions and found an apparently contradictory result: whereas frontal regions show changes in the representation of speech sounds related to comprehension, temporal regions primarily reflect the bottom-up acoustic properties of the signal (Evans and Davis, 2015). In this study, it was not possible to determine definitively what roles frontal and temporal regions play, or whether they have causal relationships in either or both directions. Lesion and stimulation studies suggest that there is a dissociation between the necessary and sufficient components of neural activity in frontal and temporal regions for speech comprehension (Pillay et al., 2017; Schomers and Pulvermüller, 2016; Zoefel and Davis, 2017), suggesting that the precise role of frontal cortex in language understanding may depend on a number of factors that were not explicitly measured here.

There is extensive evidence that representations in superior temporal regions are highly task- and context- dependent (Cibelli et al., 2015; Gagnepain et al., 2012; Holdgraf et al., 2016; Leonard and Chang, 2014; Leonard et al., 2015, 2016; Mesgarani and Chang, 2012; Sohoglu et al., 2012). This suggests that the differences observed across studies may reflect the specific parameters of each experiment, and that the cognitive and behavioral requirements of the present task primarily involved representations of speech sounds that are processed in frontal regions. It is also important to note that the regions covered by electrodes in the present study do not sample all parts of the cortical network implicated in speech comprehension (Evans et al., 2013). Both the separate and interactive roles of these regions must be considered when developing a comprehensive model of speech understanding. Further work is required to understand the exact circumstances under which representations in the various relevant brain regions change with task and stimulus parameters.

Funding: This work was supported by NIH grants F32-DC013486, R00-NS065120, DP2-OD00862, and R01-DC012379; and a Kavli Institute for Brain and Mind Innovative Research grant. E.F.C is a New York Stem Cell Foundation - Robertson Investigator. This research was supported by The New York Stem Cell Foundation, The McKnight Foundation, The Shurl and Kay Curci Foundation, and The William K Bowes Foundation.

References

Bolia, R.S., Nelson, W.T., Ericson, M.A., and Simpson, B.D. (2000). A speech corpus for multitalker communications research. J. Acoust. Soc. Am. *107*, 1065–1066.

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

Cheung, C., Hamiton, L.S., Johnson, K., and Chang, E.F. (2016). The auditory representation of speech sounds in human motor cortex. eLife *5*.

Cibelli, E.S., Leonard, M.K., Johnson, K., and Chang, E.F. (2015). The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. Brain Lang. *147*, 66–75.

Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., and Pesaran, B. (2014). Sensory-motor transformations for speech occur bilaterally. Nature *507*, 94–98.

Crone, N.E., Miglioretti, D.L., Gordon, B., and Lesser, R.P. (1998a). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain J. Neurol. *121 (Pt 12)*, 2301–2315.

Crone, N.E., Miglioretti, D.L., Gordon, B., and Lesser, R.P. (1998b). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain *121*, 2301–2315.

Dahan, D., and Mead, R.L. (2010). Context-conditioned generalization in adaptation to distorted speech. J. Exp. Psychol. Hum. Percept. Perform. *36*, 704.

Davis, M.H., and Johnsrude, I.S. (2003). Hierarchical processing in spoken language comprehension. J. Neurosci. *23*, 3423–3431.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., and Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. Neuroimage *24*, 21–33.

Edwards, E., Nagarajan, S.S., Dalal, S.S., Canolty, R.T., Kirsch, H.E., Barbaro, N.M., and Knight, R.T. (2010). Spatiotemporal imaging of cortical activation during verb generation and picture naming. NeuroImage *50*, 291–301.

Eisner, F., McGettigan, C., Faulkner, A., Rosen, S., and Scott, S.K. (2010). Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. J. Neurosci. *30*, 7179–7186.

Evans, S., and Davis, M.H. (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cereb. Cortex *25*, 4772–4788.

Evans, S., Kyong, J., Rosen, S., Golestani, N., Warren, J., McGettigan, C., Mourão-Miranda, J., Wise, R., and Scott, S. (2013). The pathways for intelligible speech: multivariate and univariate perspectives. Cereb. Cortex *24*, 2350–2361.

Gagnepain, P., Henson, R.N., and Davis, M.H. (2012). Temporal predictive codes for spoken words in auditory cortex. Curr. Biol. *22*, 615–621.

Giraud, A., Kell, C., Thierfelder, C., Sterzer, P., Russ, M., Preibisch, C., and Kleinschmidt, A. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cereb. Cortex *14*, 247–255.

Hervais-Adelman, A.G., Carlyon, R.P., Johnsrude, I.S., and Davis, M.H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. Lang. Cogn. Process. *27*, 1145–1166.

Holdgraf, C.R., De Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J.J., Knight, R.T., and Theunissen, F.E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat. Commun. *7*, 13654.

Leonard, M.K., and Chang, E.F. (2014). Dynamic speech representations in the human temporal lobe. Trends Cogn. Sci. *18*, 472–479.

Leonard, M.K., and Chang, E.F. (2016). Neural organization of speech perception: Intracranial Recording. In Neurobiology of Language, (Amsterdam: Elsevier), pp. 479–489.

Leonard, M.K., Bouchard, K.E., Tang, C., and Chang, E.F. (2015). Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex. J. Neurosci. *35*, 7203–7214.

Leonard, M.K., Baud, M.O., Sjerps, M.J., and Chang, E.F. (2016). Perceptual Restoration of Masked Speech in Human Cortex. Nat. Commun. *7*.

Liebenthal, E., Binder, J.R., Piorkowski, R.L., and Remez, R.E. (2001). Sinewave speech/nonspeech perception: An fMRI study. J. Acoust. Soc. Am. *109*, 2312–2313.

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., and Medler, D.A. (2005). Neural substrates of phonemic perception. Cereb. Cortex *15*, 1621–1631.

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

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010.

Möttönen, R., Calvert, G.A., Jääskeläinen, I.P., Matthews, P.M., Thesen, T., Tuomainen, J., and Sams, M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. Neuroimage *30*, 563–569.

Pillay, S.B., Binder, J.R., Humphries, C., Gross, W.L., and Book, D.S. (2017). Lesion localization of speech comprehension deficits in chronic aphasia. Neurology *88*, 970–975.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. U. S. A. *103*, 7865–7870.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. Science *212*, 947–949.

Schomers, M.R., and Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. Front. Hum. Neurosci. *10*.

Sohoglu, E., and Davis, M.H. (2016). Perceptual learning of degraded speech by minimizing prediction error. Proc. Natl. Acad. Sci. *113*, E1747–E1756.

Sohoglu, E., Peelle, J.E., Carlyon, R.P., and Davis, M.H. (2012). Predictive top-down integration of prior knowledge during speech perception. J. Neurosci. *32*, 8443–8453.

Wilson, S.M., Saygin, A.P., Sereno, M.I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. Nat. Neurosci. *7*, 701–702.

Zoefel, B., and Davis, M.H. (2017). Transcranial electric stimulation for the investigation of speech perception and comprehension. Lang. Cogn. Neurosci. *32*, 910–923.