Speech and speaker separation in human auditory cortex

Nima Mesgarani, Edward Chang

Department of Neurological Surgery, University of California, San Francisco, CA, USA
UCSF Center for Integrative Neuroscience, San Francisco, CA, USA
Nima.Mesgarani,changed@ucsf.edu

Abstract
Humans possess a remarkable ability to attend to a single speaker’s voice in a multi-talker background. How the auditory system manages to extract intelligible speech under such acoustically complex and adverse listening conditions is not known, and indeed, it is not clear how attended speech is internally represented. Here, using multi-electrode surface recordings from the cortex of subjects engaged in a listening task with two simultaneous speakers, we demonstrate that population responses in non-primary human auditory cortex faithfully encode critical features of attended speech: speech spectrograms reconstructed based on cortical responses to the mixture of speakers reveal salient spectral and temporal features of the attended speaker, as if listening to that speaker alone. A simple classifier trained solely on examples of single speakers can decode both attended words and speaker identity. We find that task performance is well predicted by a rapid increase in attention-modulated neural selectivity across both local single-electrode and population-level cortical responses. These findings demonstrate that the cortical representation of speech does not merely reflect the external acoustic environment, but instead gives rise to the perceptual aspects relevant for the listener’s intended goal.

Index Terms: speech synthesis, unit selection, joint costs

1. Introduction
Separating out a speaker of interest from other speakers in a noisy, crowded environment is a perceptual feat that we perform routinely. The ease with which we hear under these conditions significantly belies the intrinsic complexity of this process also known as the cocktail party problem [1,2]: Concurrent complex sounds, that are completely mixed upon entering the ear, are segregated and selected from within the auditory system. The resulting percept is that we selectively attend to the desired speaker while tuning out the others.

While previous studies have described neural correlates of masking and selective attention to speech [3], fundamental questions remain unanswered regarding the precise nature of speech representation at the juncture where competing signals are resolved. In particular, when attending to a speaker within a mixture, what key aspects (e.g. spectrotemporal profile, spoken words and speaker identity) are represented in the auditory system and how do they compare to representations of that speaker alone? How rapidly does a selective neural representation build-up when one attends to a specific speaker? Can breakdowns in these processes explain distinct perceptual failures such as the inability to hear the correct words, or follow the intended speaker?

To answer these questions, we recorded cortical activity from human subjects implanted with customized high-density multi-electrodes arrays as part of their clinical work-up for epilepsy surgery [4]. While limited to this clinical setting, these recordings provide simultaneous high spatial and temporal resolution while sampling the population neural activity from the non-primary auditory speech cortex in the posterior superior temporal lobe. We focused our analysis on high gamma (75-150Hz) local field potentials, which have been found to correlate well with the tuning of multi-unit spike recordings. In humans, the posterior superior temporal gyrus has been heavily implicated in speech perception and is anatomically defined as the lateral parabelt auditory cortex (including Brodmann areas 41, 42, and 22) [5].

2. Experiment

2.1. Methods
Three human subjects with normal hearing underwent placement of a subdural electrode array as part of their clinical treatment of epilepsy. We used speech samples from a publicly available database called Coordinate Response Measure (CRM [6]). One male and one female speaker were selected with two call signs (ringo and tiger), three colors (red, blue or green) and three numbers (two, five or seven). We generated 12 unique combinations of call sign, color and number per speaker (total of 24 single speaker phrases) and 28 mixture speech samples by selecting from combinations of the 24 single speaker sentences (0 dB Target-to-Masker ratio). Speech sounds were presented monaurally from a loud speaker. We used stimulus reconstruction [7,8] to map the population ECoG responses to the spectrogram of the speech stimulus. Reconstruction filters were estimated from neural responses to a separate speech corpus (TIMIT). Test speakers were not used in the estimation of filters. For word and speaker decoding analysis, a regularized linear classifier [9] was trained on neural responses of the single speakers and then used to decode the spoken words and speaker identity of the attended speech mixture.

2.2. Multispeaker perception task
Subjects listened to speech samples from CRM corpus [6]. A typical sentence was “ready tiger go to red two now” where tiger is the call sign, and red-two is the color-number combination. One male and one female speaker were selected, each speaking the same 12 unique combinations of two call signs (ringo or tiger), three colors (red, blue, or green), and three numbers (two, five, or seven). Example acoustic spectrograms from two individual speakers are shown in Figure 1a-b. The two voices differ along several dimensions including pitch (male vs. female), spectral profiles (different vocal track shapes) and
temporal characteristics (speaking rate). Subjects first listened to each of the speakers alone and were able to report the color and number with 100% accuracy. Subjects then listened to a monaural, simultaneous mixture of the two speakers’ phrases with different call signs, colors and numbers. The subjects were instructed to respond by indicating the color and number spoken by the talker who uttered the target call sign. The target call sign (ringo or tiger) was fixed and shown visually on a monitor during each trial block, which contained 28 different mixture sounds. Since the target speaker was changed randomly from trial to trial, the subjects were required to initially monitor both voices (divided attention) to identify the target speaker. The target call sign was switched after each block, turning the previous target speaker in each mixture into masker. This resulted in two sets of behavioral and neural responses for each identical mixture sound, which differed only in the focus of attention. Subjects reported correct responses in 74.8% of trials.

3. Results

3.1. Reconstructed speech

To determine the spectrotemporal encoding of the attended speaker, the method of stimulus reconstruction was used [6, 7] to estimate the speech spectrogram represented by the population neural responses. Reconstructed spectrograms provide an intuitive way to examine how the population neural responses encode the spectrotemporal features of speech, and more importantly, can be compared with the original acoustics as well as across attentional conditions. We first calculated the reconstruction filters from a passive listening task using a separate continuous speech corpus (TIMIT [10]) that consisted of 499 unique short sentences spoken by 402 different speakers. The filters were then fixed and applied to a novel set of population neural responses to the single and attended mixture speaker stimuli for spectrogram reconstruction.

When listening to a single speaker alone, the reconstructed spectrograms from population neural activity corresponded well to the spectrotemporal features of the original acoustic spectrograms (Fig. 1e,f compared to a,b respectively), exhibiting fairly precise temporal features and spectral selectivity (e.g. correspondence between the high frequency bursts of energy in “tiger” and “two”, in Figs. 1a,b and 1e,f). The average and standard deviation of the correlation between reconstructed and original spectrograms over 24 sentences were 0.60 ± 0.034 (0.60 and 0.62 for Fig. 1e,f examples). When attending to each of the two speakers, the reconstructed spectrograms from the same speech mixture showed a striking difference depending upon which speaker was attended (Fig. 1 g,h). For each pair, the key temporal and spectral features of the target speaker are enhanced relative to the masker speaker (Fig. 1 g,h compared to e,f respectively). To compare directly, the energy contours from these reconstructed spectrograms are overlaid in Figure 1i. Important spectrotemporal details of the attended speaker representation were extracted, while the masker speech was effectively suppressed.

Attentional modulation of the neural representation was quantified, separately for correct and error trials, by measuring the correlation of the reconstructed spectrograms from the mixture in two attended conditions with original acoustic spectrograms of the speakers alone (Fig. 2 a-d). During correct trials (Fig. 2 a,c), we observed a significant shift of average correlation values toward the target speaker representation. During error trials, in contrast, no significant shift was observed (Fig. 2 b,d). Furthermore, the correlations between reconstructed mixture and the masker speaker were higher than the average intrinsic correlation between randomly chosen original acoustic speech phrases (Fig. 2 c,d. dashed lines), revealing a weak presence of the masker speaker in mixture reconstructions, even in correct trials.
The difference in speaking rate of the two speakers, yet stereotyped structure of the carrier phrases, results in specific average temporal modulation profiles for each speaker. To investigate encoding of distinct spectral profile and characteristic temporal rhythm of the target compared to the masker speaker, we estimated the average difference between reconstructed spectrograms of the two speakers, when presented alone and in the attended mixture (Fig. 2 c,f). The comparison between the two average difference reconstructed spectrograms reveals enhanced encoding of both temporal and spectral aspect of the attended speaker.

3.2. Time course of attentional modulation

To study the time course of attention-induced modulation of reconstructed mixture spectrograms towards the attended speaker, we calculated an attentional modulation index ($AM_{\text{spec}}$), using a sliding window of 250 milliseconds throughout the trial duration:

$$AM_{\text{spec}} = \frac{\text{Corr}(SP_{\text{spec}}, SP_{\text{attend}}) - \text{Corr}(SP_{\text{spec}}, SP_{\text{2 attend}})}{\text{Corr}(SP_{\text{2 spec}}, SP_{\text{2 attend}}) - \text{Corr}(SP_{\text{2 spec}}, SP_{\text{1 attend}})}$$

(1)

where $SP_{1,2\text{spec}}$ are the original acoustic spectrograms of speakers one and two respectively and $SP_{1,2\text{attend}}$ are the spectrograms reconstructed from neural responses to the mixture with attended targets, speaker one and two, respectively. Positive values of this index reflect shifts towards the target, negative values reflect shifts to the masker representation, and values around zero reflect no shift ($AM_{\text{spec}} = 0.58$ for Fig 1 example). An upper bound for the $AM_{\text{spec}}$ was calculated by assuming that attention, at best, restores the single speaker reconstructions of the target speaker (replacing $SP_{1,2\text{attend}}$ in equation one with $SP_{1,2\text{alone}}$, Fig 2g, gray line). The $AM_{\text{spec}}$ from the mixture was first estimated from correct trials (Fig. 2g, black line), and could resolve the time point at which the reconstructed spectrograms were modulated by attention. After the end of the call sign, which cues the speaker that should be attended, a rapid positive shift in the $AM_{\text{spec}}$ was observed, implying the enhanced representation of the target speaker. In error trials, this effect shows a bias toward the masker speaker, which, in contrast, occurred far earlier in the time course. The neural response shifts towards the masker as early as the call sign suggests that listeners had prematurely attended to the wrong speaker during those error trials.

3.3. Decoding of message and speaker

While the reconstruction analyses showed clear attention-based spectrotemporal modulation, we wanted to explicitly determine whether the attended speech in a mixture could be decoded from a model of a single speaker. A regularized linear classifier [8] was trained on neural responses of the single speakers and then used to decode both the spoken words and speaker identity of the attended speech mixture. To keep the chance performance at 50% across all comparisons, classification results were limited only to the choices that were present in each mixture. For correct trials, the color and number of the attended speaker were decoded with high accuracy (77.2% and 80.2%, p<10e-4, t-test, Fig. 3a). The decoding performance during error trials however was significantly below chance (30.0%, 30.1% p<10e-4, t-test, Fig. 3b) indicating a systematic bias toward decoding the words of the masker speaker. In addition, for correct trials, the call sign was classified at chance performance (Fig. 3a). For incorrect trials however, the classifier detected the masker call sign significantly more often than the target call sign (34.1%, p<10e-4, t-test, Fig. 3b), which again shows errors due to an early selection of the masker (incorrect) speaker.

For the speaker identification analyses, we divided the behavioral error types into two subsets: The first type occurred when the reported color-number combination was incorrect for either speaker (‘incorrect’; 16.5% of trials). The second type occurred when subjects reported the correct color-number for the masker instead of the target speaker (‘correct for masker’; 8.6% of trials).

In correct trials, the classifier identified the target speaker 93.0% of the time (p<10e-4, t-test, Fig. 3c). During incorrect trials, the classifier performance was at chance. However, during correct for masker trials, the classifier identified the masker rather than the target speaker (27.3%; p<10e-4, t-test, Fig. 3c). These classification results confirm the observed restoration seen in spectrotemporal reconstruction, without necessarily assuming a linear relationship between the neural responses and the stimulus.
In summary, we demonstrate that the human auditory system powerfully restores the representation of the attended speaker while suppressing irrelevant competing speech. Speech restoration occurs at the level where neural responses still show precise phase-locking to spectrotemporal features of speech. Population responses revealed the emergent representation of speech extracted from a mixture, including the moment-by-moment allocation of attentional focus.

These results have significant implications for models of auditory scene analysis. In agreement with recent studies, the cortical representation of speech in posterior temporal lobe does not merely reflect the acoustical properties of the stimulus, but instead relates strongly to the perceived aspects of speech [4]. While the exact mechanisms are not fully known, multiple processes in addition to attention are likely to enable this high-order auditory processing, including grouping of predictable regularities in speech acoustics [11] and feature binding, and phonemic restoration. Conversely, behavioral errors appear to result from degradation of the neural representation, a direct result of inherent sensory interference such as energetic masking [12] and/or the allocation of attention.

In speech, the end-result represented in the posterior temporal lobe appears to be unaffected by perceptually irrelevant sounds, which is highly ideal for subsequent linguistic and cognitive processing. Following one speaker in the presence of another can be trivial for a normal human listener, but remains a major challenge for state-of-the-art automatic speech recognition algorithms. Understanding how the brain solves this problem may inspire efficient and generalizable solutions over current engineering approaches [13]. It will also shed light on how these processes become impaired during aging and in disorders of speech perception in real-world hearing conditions.

4. Conclusion

In summary, we demonstrate that the human auditory system powerfully restores the representation of the attended speaker while suppressing irrelevant competing speech. Speech restoration occurs at the level where neural responses still show precise phase-locking to spectrotemporal features of speech. Population responses revealed the emergent representation of speech extracted from a mixture, including the moment-by-moment allocation of attentional focus.

These results have significant implications for models of auditory scene analysis. In agreement with recent studies, the cortical representation of speech in posterior temporal lobe does not merely reflect the acoustical properties of the stimulus, but instead relates strongly to the perceived aspects of speech [4]. While the exact mechanisms are not fully known, multiple processes in addition to attention are likely to enable this high-order auditory processing, including grouping of predictable regularities in speech acoustics [11] and feature binding, and phonemic restoration. Conversely, behavioral errors appear to result from degradation of the neural representation, a direct result of inherent sensory interference such as energetic masking [12] and/or the allocation of attention.

In speech, the end-result represented in the posterior temporal lobe appears to be unaffected by perceptually irrelevant sounds, which is highly ideal for subsequent linguistic and cognitive processing. Following one speaker in the presence of another can be trivial for a normal human listener, but remains a major challenge for state-of-the-art automatic speech recognition algorithms. Understanding how the brain solves this problem may inspire efficient and generalizable solutions over current engineering approaches [13]. It will also shed light on how these processes become impaired during aging and in disorders of speech perception in real-world hearing conditions.

5. Acknowledgment

The authors would like to thank A. Ren for technical help. E.F.C. was funded by NIH grants NS065120, OD00862, DC012379, and the Ester A. and Joseph Klingenstein Foundation.

6. References