NeuroImage 117 (2015) 151–159

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Decoding speech perception from single cell activity in humans

Ori Ossmy [a,b], Itzhak Fried [c,d], Roy Mukamel [a,b,*]

[a] Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv 69978, Israel
[b] School of Psychological Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel
[c] Functional Neurosurgery Unit, Tel Aviv Medical Center and Sackler School of Medicine, Tel-Aviv University, Tel Aviv 69978, Israel
[d] Department of Neurosurgery, David Geffen School of Medicine and Semel Institute for Neuroscience, University of California at Los Angeles (UCLA), Los Angeles, CA 90095, USA

ARTICLE INFO

Article history:
Received 8 October 2014
Accepted 2 May 2015
Available online 11 May 2015

Keywords:
Single unit recording
Decoding
Speech perception
Local field potentials

ABSTRACT

Deciphering the content of continuous speech is a challenging task performed daily by the human brain. Here, we tested whether activity of single cells in auditory cortex could be used to support such a task. We recorded neural activity from auditory cortex of two neurosurgical patients while presented with a short video segment containing speech. Population spiking activity (~20 cells per patient) allowed detection of word onset and decoding the identity of perceived words with significantly high accuracy levels. Oscillation phase of local field potentials (8–12 Hz) also allowed decoding word identity although with lower accuracy levels. Our results provide evidence that the spiking activity of a relatively small population of cells in human primary auditory cortex contains significant information for classification of words in ongoing speech. Given previous evidence for overlapping neural representation during speech perception and production, this may have implications for developing brain–machine interfaces for patients with deficits in speech production.

Introduction

The ability to correctly discriminate speech is crucial for successful social interaction. To comprehend auditory content, the brain has to decipher a variety of sounds in real time. Previous electrophysiological studies in animals have successfully used spiking activity in auditory cortex to classify different sounds including species-specific vocalizations (e.g., grasshoppers (Machens et al., 2003); song birds (Grace et al., 2003; Narayan et al., 2006); cats (Gehr et al., 2000); monkeys (Russ et al., 2008)), or vocalizations across species (e.g., marmoset calls in ferrets (Schnupp et al., 2006); marmoset calls in cat (Wang and Kadia, 2001); bird chirps in cats (Chechik et al., 2006)).

In humans, discrimination of speech content has been demonstrated using various non-invasive techniques. Functional magnetic resonance imaging (fMRI) studies showed cortical representation of speech based on spatial activation patterns in Heschl's gyrus (Formisano et al., 2008; Wessinger et al., 2001; Binder et al., 2000). Other studies using Magnetoencephalography (MEG) found that the degree of correspondence between the temporal envelope of the signal in auditory cortex and stimulus soundwave co-varies with the level of speech comprehension (Ahissar et al., 2001). Furthermore, it has been found that the phase of the MEG signal in the theta-band (4–8 Hz) reliably discriminates spoken sentences (Luo and Poeppel, 2007).

Invasive studies using Electrocorticography (ECoG) have shown that cortical responses in the superior temporal gyrus (STG) track the envelope of attended speech streams (Zion Golumbic et al., 2013; Mesgarani and Chang, 2012; Canolty et al., 2007). Others found that the STG is robustly organized according to sensitivity to basic phonetic items (Mesgarani et al., 2014; Chang et al., 2010) and that slow and intermediate temporal fluctuations corresponding to syllable rate can be reconstructed based on power in high-gamma frequency band (Pasley et al., 2012). It has also been shown that the ECoG signal from electrodes implanted in Heschl's gyrus (HG) follows the temporal speech envelope over a wide range of speaking rates (Nourski et al., 2009) and can be used to facilitate discrimination of voiced from unvoiced phonemes (Steinschneider et al., 2005). Despite this comprehensive research, the relative contribution of spiking activity and optimal features of the rich LFP signal in auditory cortex in decoding perceived words from ongoing speech is not known.

It has been previously shown that activity in auditory cortex during passive perception overlaps with activity during overt (Zheng et al., 2010; Flinker et al., 2011; Cogan et al., 2014) and covert speech (Buchsbaum et al., 2001; Pei et al., 2011; Martin et al., 2014). Under these circumstances, characterizing the activity patterns of single cells during passive perception may also have important implications for comprehending the process of speech production (Bouchard et al., 2013).

In the current study, we recorded spiking activity and local field potentials from the putative primary auditory cortex of two neurosurgical patients while they were presented with an audio–visual stimulus containing on-going speech monologue. We used a support vector machine

* Corresponding author at: School of Psychology, Tel-Aviv University, Ramat-Aviv 69978, Israel.
E-mail address: rmukamel@tau.ac.il (R. Mukamel).

http://dx.doi.org/10.1016/j.neuroimage.2015.05.001
1053-8119/

(SVM) classifier in order to discriminate 6 different words and detect their onset using information from spiking activity. We also examined local field potentials (LFPs) and found that across various features, phase in the low frequency band (8–12 Hz) was best for decoding words, although performance was much lower compared with using population spiking activity. Combining information from spikes and low frequency LFP phase improved classification performance compared to using data from either signal alone.

Materials and methods

Patients and electrophysiological recording

Data was collected from two patients (21 years old male and 19 years old female) with pharmacologically intractable epilepsy, implanted with intracranial depth electrodes to identify seizure focus for potential surgical treatment (Mukamel and Fried, 2012). Electrode location was based solely on clinical criteria. Each electrode terminated in a set of nine 40-μm platinum–iridium microwires (Fried et al., 1999) — eight active recording wires, referenced to the ninth. Signals from these microwires were recorded at 28 kHz for the first patient and 30 kHz for the second patient using a 64-channel acquisition system. Before surgery each patient underwent placement of a stereotactic headframe, and then a detailed MR image was obtained using a spoiled-gradient sequence, followed by cerebral angiography. Both anatomical and angiography images were transmitted to a workstation in the operating room, and surgical planning was then performed, with selection of appropriate temporal and extra-temporal targets and appropriate trajectories based on clinical criteria. To verify electrode position, CT scans following electrode implantation were co-registered to the preoperative MRI using Vitrea® (Vital Images Inc.). The patients provided written informed consent to participate in the experiments. The study was approved by and conformed to the guidelines of the Medical Institutional Review Board at UCLA. Data collected from the first patient was previously reported (Mukamel et al., 2011; Bitterman et al., 2008; Nir et al., 2007).

Stimuli and behavioral task

Patients observed nine repetitions of a 17 s long audio–visual clip at their bedside. The clip was taken from the movie "The Good, The Bad, and The Ugly" (starting from minutes 44:31 in the original film) and is comprised mainly of speech monologue containing 23 words and environmental sounds. The patients' task was to follow the plot.

Data preprocessing

To detect spiking activity, the data was band-pass filtered offline between 300 and 3000 Hz and spike sorting was performed using WaveClus (Quiroga et al., 2004), similar to previous publications (Quiroga et al., 2005). This process yields for each detected neuron a vector of time stamps (1 ms resolution) during which spikes occurred.

We assessed whether the spiking activity of the recorded neurons is evoked by different spoken words — 'Now', 'Tight', 'Right', 'Neck', 'Pig' and 'Rope', embedded in the speech sequence. These words were chosen since they fit within a time window of 250 ms without overlapping with adjacent words. The spike train of each neuron during the 250 ms time window aligned to specific word onset was extracted and spike counts were calculated in twenty, 12.5 ms consecutive time bins. In order to assess responsiveness of each neuron to the various stimuli, we examined the degree of repeated spike patterns across trials. To this end the binned signals were averaged across odd and even trials separately and the Pearson correlation coefficient between the two averages was computed. Cells exhibiting correlation coefficients greater than 0.45 (lowest statistically significant correlation level when using 20 bins) for at least one word were considered responsive and taken for further analysis.

Word classification

We used a multi-class support vector machine to discriminate among the six different words within the speech sequence. We used a Matlab implementation of a SVM classifier (Chang and Lin, 2011; software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm) and least squares as a cost function. Accuracy levels were compared with a null distribution obtained by shuffling the labels of the data and performing the same classification procedure as on the original data.

Spiking activity from time windows corresponding to individual words was binned in consecutive non-overlapping temporal windows. Thus, the data of each word consisted of 9 matrices (one for each trial). The value in each matrix cell *i,j* corresponded to the spike count of neuron *i*, in time bin *j*. During each classification iteration we performed a standard "leave-one-out" procedure in which one matrix of each of the six words was randomly chosen as test data and the classifier was trained to discriminate the 6 words based on the remaining matrices. During the test stage, the classifier assigned labels to left-out matrices (the trials which it was not trained on) and its performance was assessed. This procedure was iterated 500 times.

We estimated the optimal temporal resolution for classification by varying the size of non-overlapping bins. Performance level of word classification was assessed using different bin sizes as input to the classifier (either 25 ms, 50 ms, 125 ms, or 250 ms; corresponding to 10, 5, 2, and 1 temporal bins respectively). Thus given N neurons, the population spike response representation of one word during one trial using, for example, 50 ms bins is an $N \times 5$ matrix of spike counts.

Detection of word onset

We also assessed whether we can detect the correct time segments (250 ms) of each of the six word instances within the complete ongoing 17 s long audio–visual segment. We trained a binary classifier to discriminate between word and non-word bins (see below) in order to detect word onset. First, we set aside data from one trial (number of neurons $\times$17,000 ms long population spike train) to be used later as test set. For each word, we extracted 250 ms spike trains corresponding to word onset from the remaining eight trials. These spike trains were binned by calculating the spike count in five consecutive 50 ms temporal windows resulting in eight matrices (one for each trial; matrix size = number of neurons $\times$5) which were labeled 'word' bins. The same process was performed with a randomly chosen time point within the 17-s long sequence. This resulted in another eight matrices which were labeled as 'non-word' bins. These two sets of eight labeled matrices were used to train a classifier to discriminate 'word' from 'non-word' bins.

Next, we took the 17-s spike train that was set aside. Spiking activity from the first time window of 250 ms was taken and binned to five consecutive 50 ms bins (similar to the procedure performed with the training data). This matrix (number of neurons $\times$5) was used as test data to the classifier which labeled it as either belonging to 'word' or 'non-word' bin (based on the mapping rule learned from the training data). In this manner, the classification procedure yielded a label for each time bin. This process was iterated in 10 ms increments (i.e., classifying spike trains from time 10–260 ms in the following step and so on until the final time bin 16,750–17,000 ms). This resulted in a vector (length = 1676) of 'word'/'non-word' labels.

The entire process was iterated 500 times (each time using a different randomly chosen time point to be used as 'non-word' bins during training) and the percentage of 'word bin' labels assigned for each time window across iterations was calculated. The window with the maximal percentage was assigned as the classified time window of word onset. We performed this analysis for each word separately