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

resulting in 6 time windows corresponding to the decoded onset of each individual word. These time windows were considered correct if their onset was within the limits of 50 ms before and after the real word onset. Otherwise, it was considered as a false alarm. We had 9 trials and 6 words therefore a maximum of 54 correct detections across all classifications. Detection accuracy was determined by calculating the mean percentage of hits in decoding onset of all words across all test trials. Chance performance was assessed by repeating this procedure using randomly shuffled labels of the training data in each classification iteration.

Local field potentials (LFPs)

After assessing classification performance using spiking activity, we examined decoding of word identity using various features of the LFP signal. First, we determined which aspects of this rich signal could be used as input to the decoder. To that end, the raw signal was first notch filtered to remove 60 Hz electrical noise using a 2nd order Butterworth filter between 59 and 61 Hz (implemented by using Matlab's 'filtfilt' function that results in zero phase shifting). Next, the signal was downsampled from the original (28 kHz for first patient or 30 kHz for second patient) to 1 kHz using MATLAB *resample* function.

Next, we used LFP channels from which spiking activity of neurons was detected (13 and 5 channels for first and second patient respectively) and examined which features of the LFP signal are evoked by the stimulus. The features we examined included phase, and power, in all frequencies up to 120 Hz. For each word, we took the corresponding 250 ms time window, and used the Hilbert transform to calculate the phase and power of the LFP signal across all frequencies. Signal features that are evoked by the word stimulus will be coherent across presentation trials while signal features that are not, are expected to exhibit low coherence values across trials. For each word and each LFP channel, we computed the coherence in phase or power between trials as a function of frequency using the formulas below:

$$\boldsymbol{CPhase_i} = \left(\frac{\sum_{n=1}^{N}(\cos\boldsymbol{\theta_{ni}})}{\boldsymbol{N}}\right)^2 + \left(\frac{\sum_{n=1}^{N}(\sin\boldsymbol{\theta_{ni}})}{\boldsymbol{N}}\right)^2$$

$$\boldsymbol{CPower_i} = \frac{\sqrt{\frac{\sum_{n}^{N}\left(\boldsymbol{A_{ni}^2} - \overline{\boldsymbol{A}}_{\boldsymbol{ni}}^2\right)^2}{\boldsymbol{N}}}}{\overline{\boldsymbol{A}}_{\boldsymbol{i}}^2}$$

where N is the number of trials (in our case N = 9), and θni and Ani are the phase and amplitude at frequency i, and trial n. The above functions provide a measure of the coherence between trials for each frequency. Both measures increase as a function of similarity across trials. For each LFP channel we had six such coherence functions for phase and six coherence functions for power (one for each individual word).

Finally, we calculated a dissimilarity index to examine the signal features that are not only evoked by the stimulus but also selective to particular words. To that end, the coherence across trials of a particular word was compared with the coherence across 9 randomly chosen trials of the remaining words. A dissimilarity function between the coherence of trials of the same word and coherence of trials of different words was computed across all frequencies:

$$\boldsymbol{Dissimilarity\_Phase_i} = \boldsymbol{Cphase_{i,same}} - \boldsymbol{Cphase_{i,different}}$$
$$\boldsymbol{Dissimilarity\_Power_i} = \boldsymbol{Cpower_{i,same}} - \boldsymbol{Cpower_{i,different}}$$

where i corresponds to a particular frequency, and same/different refers to trials of the same or different words respectively. These dissimilarity functions were computed for each individual word and averaged across words for each LFP channel. The average function (dissimilarity vs. frequency) of each LFP channel was then Z-score normalized (i.e., remove mean and divide by SD) and averaged across channels.

This index is a measure of the degree of signal similarity between trials of the same word relative to the degree of signal similarity between trials of different words. Two criteria need to be fulfilled in order to get significant positive dissimilarity index values: 1) robustness — the signal must be consistent across trials of a given word (strong coherence between trials of the same word) and 2) selectivity — the signal must have different coherence values across different words. If the signal is inconsistent across trials of the same word or consistent across trials but not different across different words, dissimilarity index values will be low. If, on the other hand, the signal is similar across trials of the same word and distinct across all the other words, dissimilarity index values will be high. High dissimilarity index value of a particular LFP signal feature implies that this LFP signal feature is robust and selective — therefore a good candidate to be used for decoding. After establishing the most robust and selective LFP feature across words, it was used in a multi-class SVM classifier to perform a similar classification procedure which was performed using spiking activity (see Word classification).

Results

We recorded spiking activity and LFPs from two patients while presented with an audio–visual segment containing speech (see Materials and methods). Patients were implanted bilaterally with depth electrodes in auditory cortex for clinical purposes (see Fig. 1). In patient 1, the left Heschl's gyrus (HG) was bifurcated and the electrode was on the posterior bank, in the middle of the medial/lateral axis. The right HG of this patient was trifurcated and the electrode was on the most anterior portion, and in the middle of the medial/lateral axis. These regions are slightly anterior to where primary auditory cortex is typically located (Rademacher et al., 2001). For additional anatomical details on the exact placement of electrodes in this patient see description of patient 2 in Mukamel et al. (2011). In patient 2, the left HG was bifurcated and the electrode was also in the most posterior portion. Her right HG was trifurcated and the electrode was on the posterior bank. Overall across the two patients we found extra-cellular spiking activity from 43 neurons, and obtained LFPs from 18 different channels (see Table 1).

The patients were presented with nine repetitions of a 17-s long audio–visual excerpt containing speech. We assessed whether the spiking activity of the recorded neurons is evoked by the different spoken words ('Now', 'Tight', 'Right', 'Neck', 'Pig' and 'Rope'; see Materials and methods). Fig. 2 demonstrates the spiking activity (raster plots) of one neuron in four different time frames corresponding to four words. As can be seen by the repeatable pattern across trials, this neuron responded strongly and robustly to all four words. The total number of spikes evoked by the population of neurons within the 250 ms temporal window was not significantly different across the 6 words (repeated measures ANOVA across all neurons, average p across patients = 0.98). The average firing rates during the complete audio–visual sequence across neurons was 2.17 ± 2.8 Hz (mean ± SE).

Next, we assessed whether the information conveyed by the temporal dynamics of the population spiking activity could be used to discriminate among the six different words using a support vector machine classifier. As input to the classifier, we used a matrix representing the spike counts of the population in non-overlapping temporal bins (using 4 different bin sizes — either 25 ms, 50 ms, 125 ms, or 250 ms; see Materials and methods). Fig. 3 shows classification performance as a function of bin size in each one of the patients. Significant, above-chance classification performance was obtained using bin sizes of 50 ms and 25 ms (mean accuracy across patients: using 50 ms bins = 66%, shuffled data accuracy = 17.3%, $p < 10^{-13}$ two-tailed paired *t*-test compared to shuffle, collapsed across patients; accuracy using 25 ms bins = 61%, shuffled data accuracy = 16.5%, $p < 10^{-12}$). Between the two significant bin sizes, accuracy level obtained with 50 ms bins yielded significantly

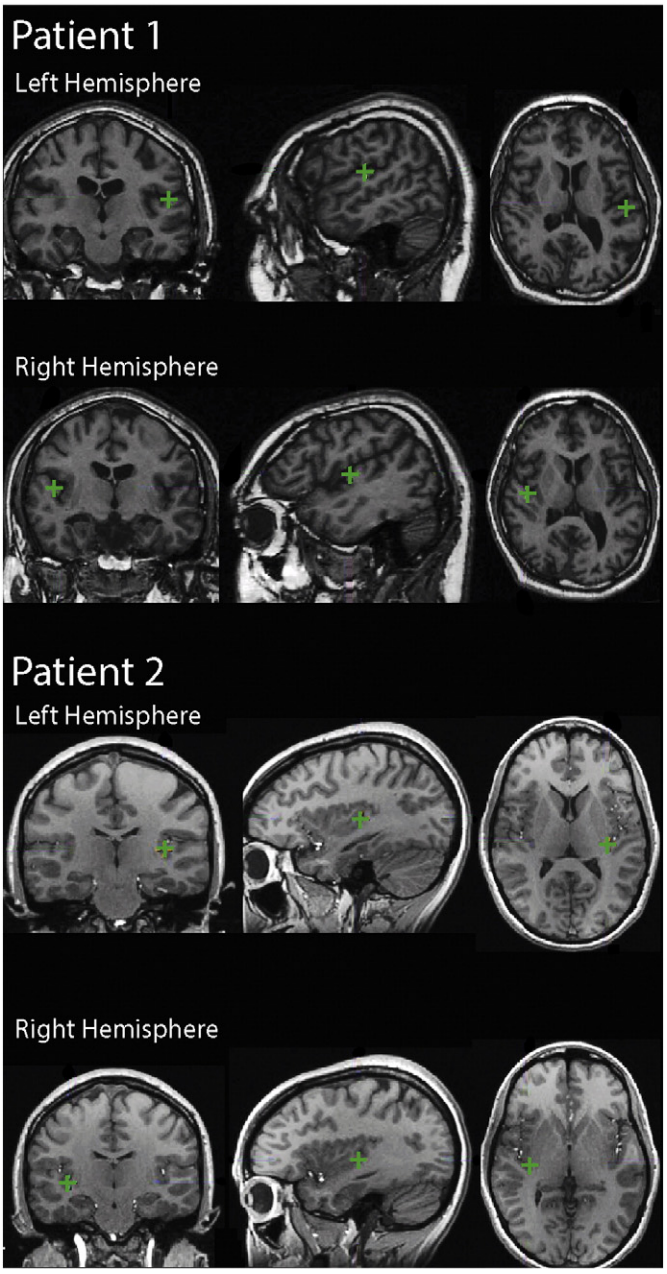

Fig. 1. Anatomical localization. Electrode locations (green dots) displayed on coronal, sagittal and axial MRI slices for patient 1 (top) and patient 2 (bottom). The electrodes in both patients were located in the posterior portion of Heschl's gyrus near putative primary auditory cortex.

higher classification levels compared with 25 ms (paired *t*-test, $p < 0.002$, collapsed across patients and words). Classification performance using larger bins (125 ms and 250 ms) yielded performance levels not significantly different from chance (accuracy using 125 ms bins: real data = 19.1%, shuffled data = 16.9%; accuracy using 250 ms bins: real data = 17%, shuffled data = 16.2%; $p = 0.31$ for 125 ms bins and $p = 0.74$ for 250 ms bins, two-tailed paired *t*-test, collapsed across patients and words). These results suggest that the temporal dynamics of population spikes during word perception contain information that is specific to word identity that is lost when simply using the total spike count.

Table 1
Recording details. Distribution of recorded spiking activity and LFP channels across hemispheres and patients.

	Left hemisphere	Right hemisphere
Patient 1	Spiking activity: 3 single units, 2 multi units. LFP: 5 channels.	Spiking activity: 9 single units, 8 multi units. LFP: 8 channels
Patient 2	Spiking activity: 11 single units, 10 multi units. LFP: 5 channels	None

Two of the 6 words ('now' and 'rope') repeated twice throughout the speech sequence thus we had two instances of these words. These two instances were semantically similar (i.e., the same word) but not identical in terms of their soundwave. The maximal cross-correlation between the absolute value of the sound wave of the first instance and that of the second instance was 0.63 averaged across both 'now' and 'rope'. For comparison, maximal cross-correlation between the sound wave of their first instance and that of other words was 0.53 averaged across both repeated words. Fig. 4 shows the 8-way confusion matrix for decoding word identity — this time including the two repeats of these words. The diagonal represents correctly classified words, and the off-diagonal represents the distribution of misclassified trials across the remaining words. The classifier reached mean performance level of 57.3% for patient 1 and 50.6% for patient 2 ($p < 5^{\circ}10^{-9}$ two-tailed paired *t*-test compared to shuffled data accuracy = 12.9% averaged across patients; theoretical chance level = 12.5%). However, this is an underestimate since the two words ('now' and 'rope') were commonly misclassified across the two instances (see 'Now'\'Now2' and 'Rope'\'Rope2' in Fig. 4). Performing a binary classification of the two instances did not achieve significance (averaged performance across patients for the word 'Rope' = 51.8% and for the word 'Now' = 55.4%; chance = 50%). Since visual input and low-level sound features were different for the two instances, this supports the notion that the evoked neural responses correspond with the auditory content. Indeed, when considering the misclassification within the same words as correct, accuracy levels rise to 76.5% in patient 1 and 67.5% in patient 2.

We examined whether the high performance we observed in decoding word identity is unique to word stimuli. To that end, we randomly extracted eight 250 ms time windows from the full audio–visual sequence containing non-speech environmental sounds. Then we conducted a similar analysis (using 50 ms bins) to classify neural activity as belonging to one of the 6 different temporal windows. Although classification performance was significantly above chance level ($27.8 \pm 8.3\%$; mean ± SD across 500 iterations; $p < 0.05$ paired *t*-test compared to shuffled results = $13.1 \pm 1.9\%$), environmental sounds classification performance in both patients was significantly lower than word stimuli classification performance ($p < 0.01$; paired *t*-test).

We also used SVM-based decoders to detect 6 different time bins that represent the correct time segments of the words within the complete spike train of each one of the trials (see Materials and methods). We found that the population activity from the recorded cells is sufficient to detect word onsets with an accuracy level of 42.6% averaged across trials. Detection accuracy in the case of shuffled data reached 5.5% ($p = 0.01$ paired *t*-test; see Materials and methods).

Next, we examined whether the LFP signal is also stimulus driven and could be used for word classification. Based on channels from which spiking activity of neurons was detected (13 channels in patient 1 and 5 channels in patient 2; see Table 1), we calculated dissimilarity index functions for LFP phase and power in all frequencies (see Fig. 5a). The graph in Fig. 5b represents the average normalized Dissimilarity index function across channels for each patient. In the case of phase, the average index across channels was significantly greater than zero in low frequencies (7.81–11.7 Hz for both patients and also 19.5 Hz in patient 2; *t*-test across all channels, $p < 0.05$ Bonferroni corrected for multiple comparisons across 59 frequencies; Fig. 5b left panel). In the case of LFP power, the average dissimilarity index across channels was not significantly different than zero across all frequency bands (*t*-test across all channels, $p > 0.05$ Bonferroni corrected for multiple comparisons across 59 frequencies; Fig. 5b right panel).

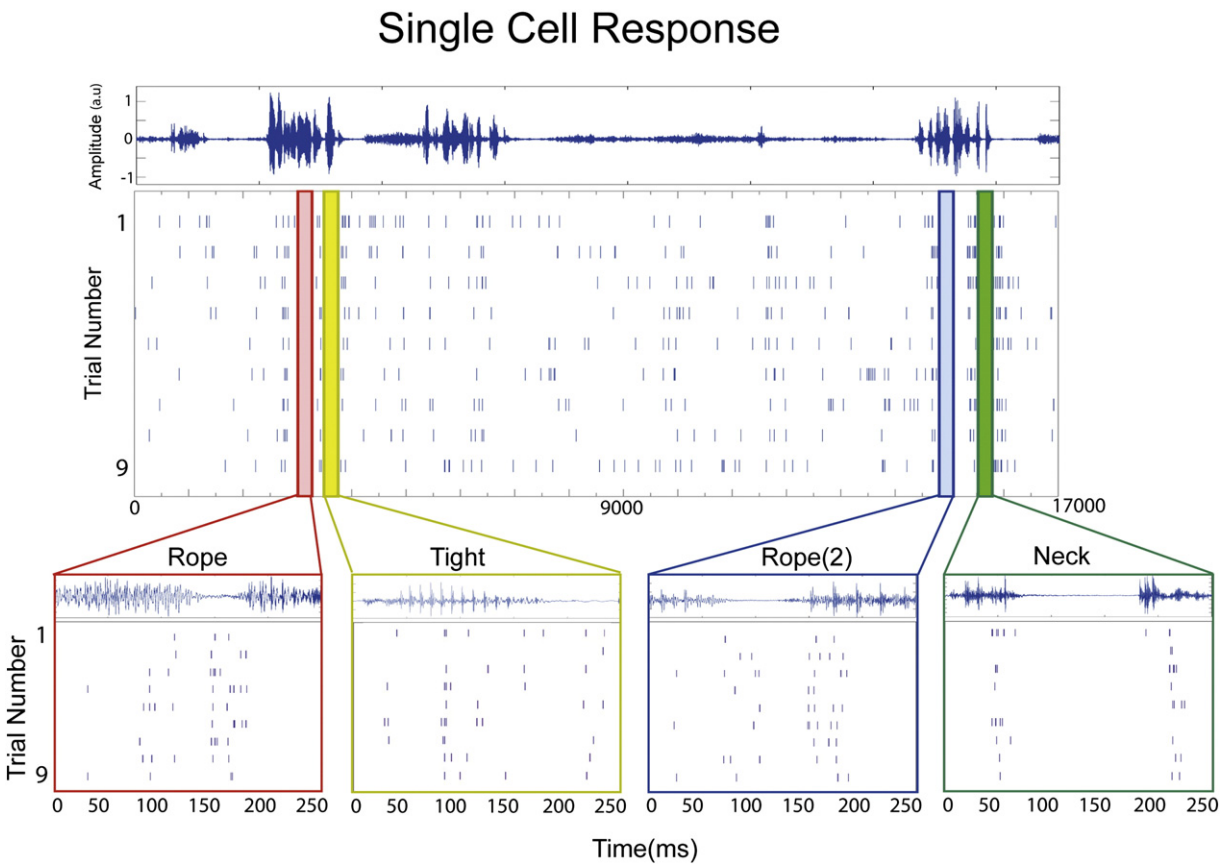

Fig. 2. Evoked spiking activity. Soundwave (top) and corresponding raster plots (middle) of a single neuron during the full speech sequence and during 4 different time-points (bottom panel) — corresponding to the words 'Tight', 'Neck' and two instances of the word 'Rope'. Full sequence duration was 17 s and word duration was 250 ms (x axis).

After establishing that phase in the low frequencies is the most robust and selective LFP feature across words, we calculated LFP phase in the 8–12 Hz frequency range in the time bins corresponding to the 6 different words and used an SVM classifier to perform a similar classification to what we performed using spiking activity. The average performance across words reached 47.8% in patient 1 ($p = 0.01$, two-tailed paired *t*-test compare to shuffled data accuracy = 16.9%) and 41.1% in patient 2 ($p = 0.02$; shuffled data accuracy = 17.1%; theoretical chance level for 6-way classification = 16.6%). Next, we added the two additional instances of the words 'now' and 'rope' and performed 8-word classification. The classifier significantly labeled words with 33% accuracy (average across both patients and 8 words; $p < 10^{-5}$ two-tailed paired *t*-test compare to shuffled data accuracy = 12.8%; theoretical chance level for 8-way classification = 12.5%). Fig. 6 displays the 8-word confusion matrix of classification level across words for each patient. Similar to the accuracy using spiking activity, this is an underestimate since the two repeated words were commonly misclassified across the two instances. Considering such misclassifications as correct, accuracy levels rise to 47.8% in patient 1 and 41.1% in patient 2. Performing a 2-way classification between the two repeats did not achieve statistical significance ('Rope' = 56.1%, 'Now' = 52.2%; averaged accuracy across patients), supporting similar evoked phase in the 8–12 Hz frequency range between similar words in different temporal positions within the speech sequence.

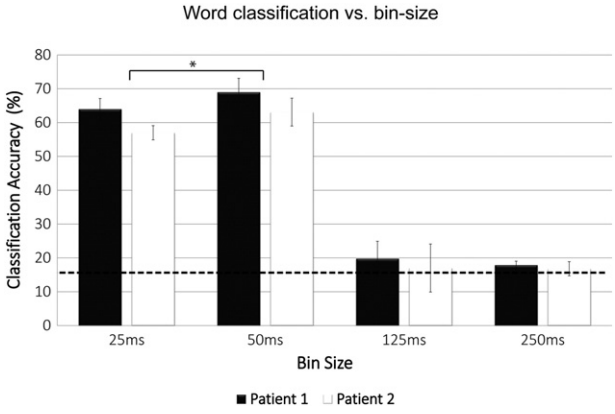

Fig. 3. Average word classification performance as a function of bin size. Spike counts from all responsive neurons were calculated using various bin sizes (x axis) and classification performance across the six words was assessed, for the two patients (see Materials and methods). 6-word classification performance reached 69% in patient 1 and 63.1% in patient 2 when using 50 ms bins; theoretical chancel level = 16.6% (black dashed line). Asterisk denotes significance of the accuracy using 50 ms bins compare to using 25 ms bins, and error bars represent S.E. (see Results).

Decoding performance using spiking activity was significantly higher compared to using low-frequency LFP phase (6-word decoding: $p = 0.04$ for patient 1 and $p = 0.02$ for patient 2; 8-word decoding: $p = 8^{\circ}10^{-3}$ for patient 1 and $p = 0.016$ for patient 2; two-tailed paired *t*-test). This result indicates that the information conveyed by the spiking activity is better suited for word discrimination compared with the optimal feature (with the highest dissimilarity index value) of the LFP signal.

Finally, we examined whether using both types of information (the population spiking activity and the LFP phase) improves word classification performance. For a given word, the input to the classifier was either a vector representing the temporal modulation of the spike count (five 50 ms bins) or a scalar representing the low-frequency LFP phase (8–12 Hz). The output of the classifier was a probability matrix representing the estimated probability that a given trial belongs to

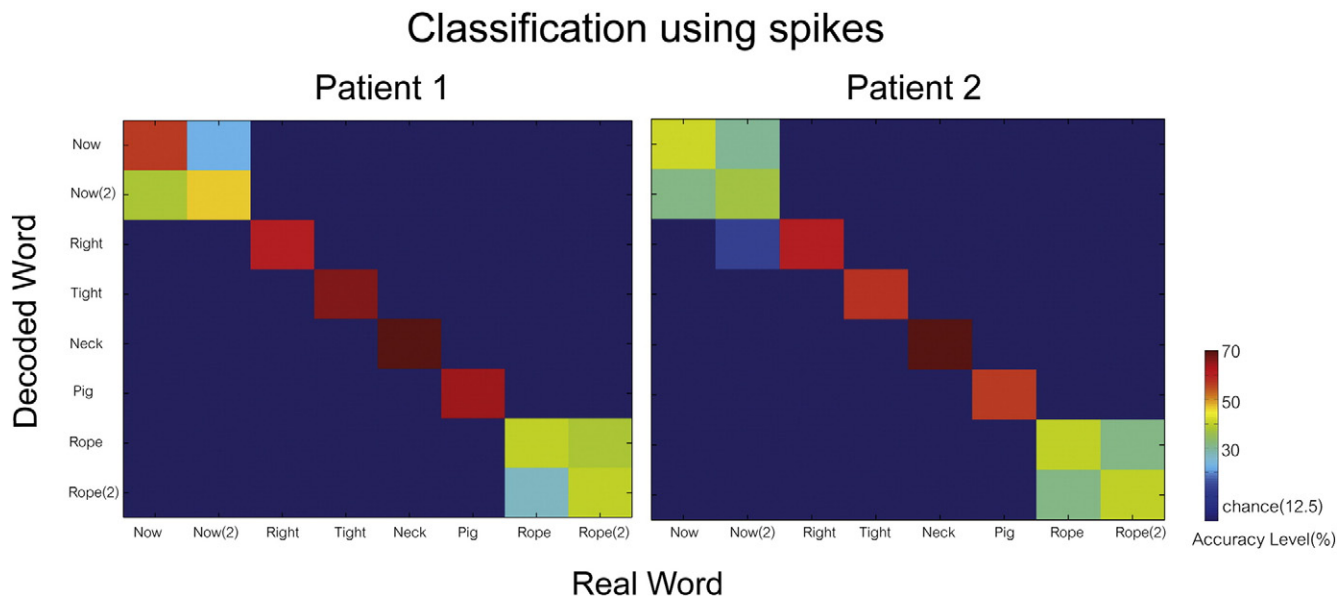

Fig. 4. Word classification performance using spikes. Confusion matrix of the 1st (left) and 2nd (right) patient. Each bin in the matrix corresponds to the proportion of iterations in which data belonging to a specific word (x axis) was decoded as belonging to one of the other words (y-axis). We had 4 different words that appeared once and 2 words that appeared twice throughout the 17-long audio–visual segment. The diagonal represents the proportion of correctly labeled words. Theoretical chance level (12.5%) is represented in blue.

any of the different words. We obtained two probability matrices (one based on spikes, and one based on LFP phase in the 8–12 Hz range). The two probability matrices were multiplied, and the class with the highest joint probability was declared as the label of the test-trial. As described earlier, classification performance based on spikes was much higher compared with that based on LFP phase. However, classification performance using spikes improved when information from LFP phase was added. This improvement in 8-word classification performance was significant for patient 1 (from 57.3% to 66.7%, $p = 0.039$) and near-significant in patient 2 (from 50.6% to 54.1%, $p = 0.058$; one-tailed paired *t*-test). We also examined changes in classification performance by increasing the dimensionality of the input to the classifier and providing both binned spiking activity and low frequency phase signals directly to the same SVM classifier. This procedure yielded similar results as when multiplying the probability matrices of two SVM classifiers as reported above.

Discussion

In the current study, we examined decoding of single words from ongoing speech using information from spiking activity and local field potentials in human auditory cortex. We find that spike counts in 50 ms bins across small population of neurons (~20) carry sufficient information to discriminate words during continuous speech perception and identify word onset. We also find that the aspect of the local field potentials signal that demonstrated the highest discriminability between words was phase in the low frequency range (8–12 Hz). LFP power, on the other hand, across all frequency ranges was more variable. The use of low frequency phase from LFP channels allowed classification of perceived words with significant accuracy levels although accuracy levels were much lower than those achieved using population spiking activity. Finally, combining data from both spiking activity and LFP phase improved classification performance relative to using data from either signal alone, although this issue deserves further investigation.

Previous studies using spike trains to discriminate auditory stimuli, have pointed to a short time scale (on the order of ~10–50 ms), representing the optimal resolution at which evoked spikes transmit information. For example, it has been reported that the optimal resolution at which neurons in ferret primary auditory cortex discriminate marmoset calls is about 40 ms (Schnupp et al., 2006). Similarly, it has been reported that the optimal resolution for discriminating bird songs in avian area L is ~10 ms (Narayan et al., 2006; Wang and Kadia, 2001). In monkey superior temporal gyrus, reported time scales are ~20 ms for conspecific vocalization (Russ et al., 2008).

In human auditory cortex, distinct time scales on the same order have been proposed for speech (Poeppel, 2003) and non-speech stimuli (Boemio et al., 2005). Compatible with this notion, our current data demonstrates that the optimal resolution at which cell populations in human auditory cortex convey information regarding word identity is 50 ms (Fig. 3). However, the information available for our decoding algorithm stems from only ~20 cells per patient. It has been previously shown that decoding accuracy can improve as a function of the number of recorded cells (Chan et al., 2014; Fried et al., 2011). Nevertheless, the fact that we could discriminate words within a long speech sequence using only ~20 cells and detect their onset implies that at increasing population size, spike count modulations can carry sufficient information for classifying auditory content. Furthermore, using environmental sounds (rather than words) as input to the classifier yielded reduced performance, suggesting that the neurons did not respond to simple spectral properties.

In terms of the local field potentials, the current findings are consistent with previous human studies using MEG (Ahissar et al., 2001; Luo and Poeppel, 2007; Patel and Balaban, 2000) and LFP recordings in primate auditory cortex (Kayser et al., 2009) demonstrating phase in the low frequencies as most informative regarding auditory stimulus identity (Lakatos et al., 2007). As to LFP power changes, intracranial studies in humans showed that speech perception evokes increased power in high gamma frequencies that are sometimes accompanied by decreased power in lower frequencies (Pasley et al., 2012; Young, 2008; Joris et al., 2004; Hickok and Poeppel, 2007). However, in the current study LFP power did not allow significant word discrimination, similar to a previous intracranial study that did not find word-specific LFP responses during speech perception and production (Chan et al., 2014).

Classification performance using spikes was always better compared with LFP phase. Combining the information from spikes and LFPs could in principle influence classification performance in two ways depending on the type of information contained in each signal. If there is significant redundancy in information content, classification performance cannot improve (in fact, it could even deteriorate since the dimensionality of

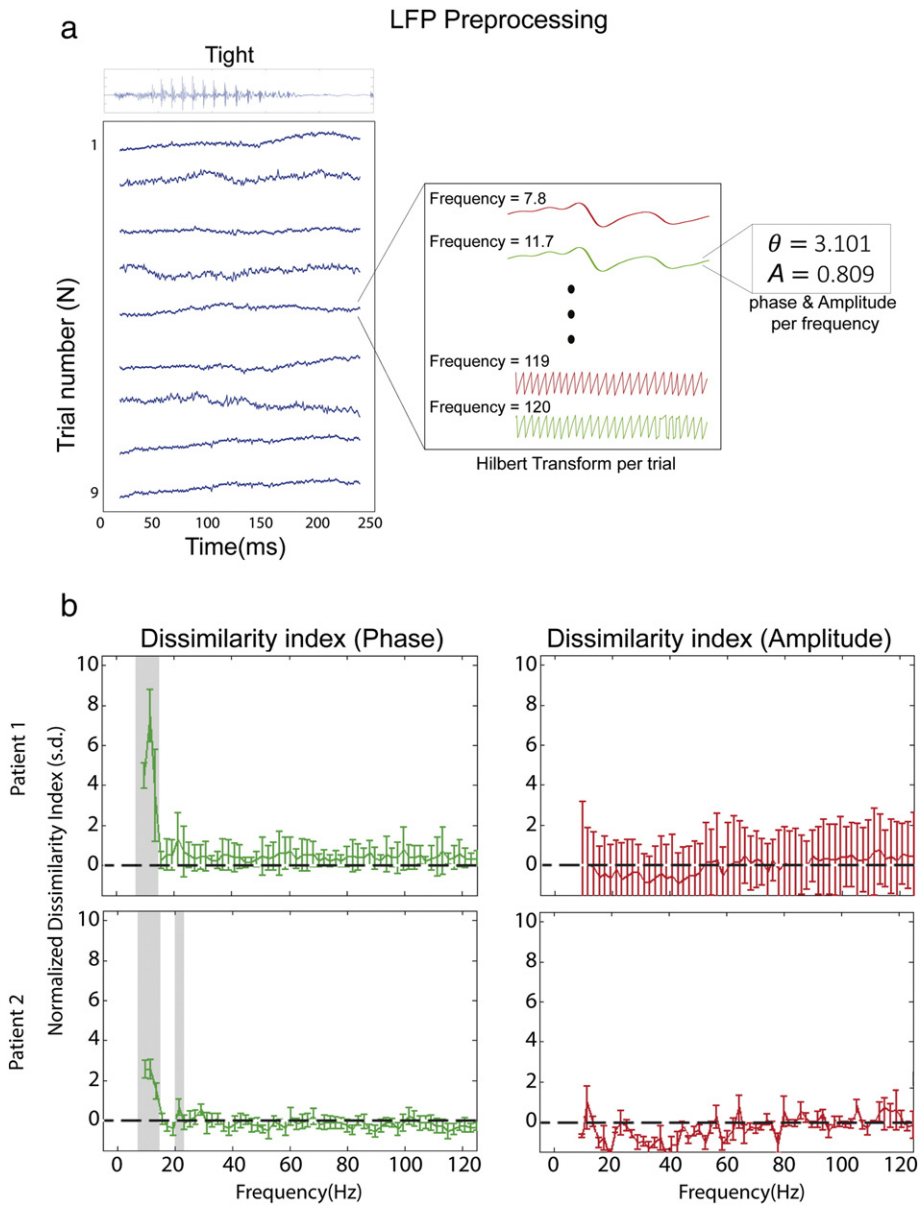

Fig. 5. Local field potentials (LFP). (a) LFP preprocessing — raw data (250 ms windows from word onset) were extracted for each word. Left panel demonstrates the signal from one channel across all trials corresponding to the word 'Tight'. The signal in each trial was decomposed using Hilbert transform to the different frequencies (right panel demonstrates decomposition of the 5th trial). The phase and amplitude were calculated per frequency. (b) Coherence across trials was calculated separately for LFP Phase and Power. We compared the coherence of the LFP signal (either phase or amplitude at different frequencies) across different trials from the same word vs. the coherence between trials taken from different words. This was conducted for all frequencies up to 120 Hz (see Materials and methods). The phase of the low frequencies (8–12 Hz) had the highest dissimilarity index across trials. Gray background denotes frequencies in which the dissimilarity index was significantly above zero (dashed black line).

the input signal to the classifier only increases). If on the other hand, the information contained in LFP phase is not available in the spiking data, classification performance can improve beyond that obtained when using information from spikes alone. We found that classification performance increased when combining the information from spike counts and LFP phase. This suggests that the two signals contain complementary information. Nevertheless, this improvement (which was found significant in the first patient but only nearly-significant in the second) requires further examination in the future.

Taken together, our data support the use of population spiking activity for decoding heard and spoken language (Creutzfeldt et al., 1989; Tankus et al., 2012). These results have implications for attempts to reveal the content of speech using brain–machine interfaces (BMIs). The recently emerging field of BMIs involving speech (Brumberg et al., 2010) requires decoding intended speech from output of cortical cells. It is thus important to know the relevant physiological features containing the highest amount of information for decoding speech content. Previous studies (Guenther et al., 2009) predicted intended speech directly from the activity of neurons in the motor cortex during overt speech production. Other studies report that ECoG activity can be used to decode vowels and consonants in spoken or imagined words (Martin et al., 2014; Pei et al., 2011). It was found that within auditory cortex, decoding accuracies were similar for overt and covert speech. In contrast, activity in motor areas only allowed discrimination during overt speech. This implies that covert speech processing consists mainly of imagining what the word sounds like, rather than simulating the motor actions necessary for speech production. Together with our data, showing that speech perception can be decoded successfully from spiking activity in auditory cortex, it remains to be seen whether single unit activity in auditory cortex during covert speech can be

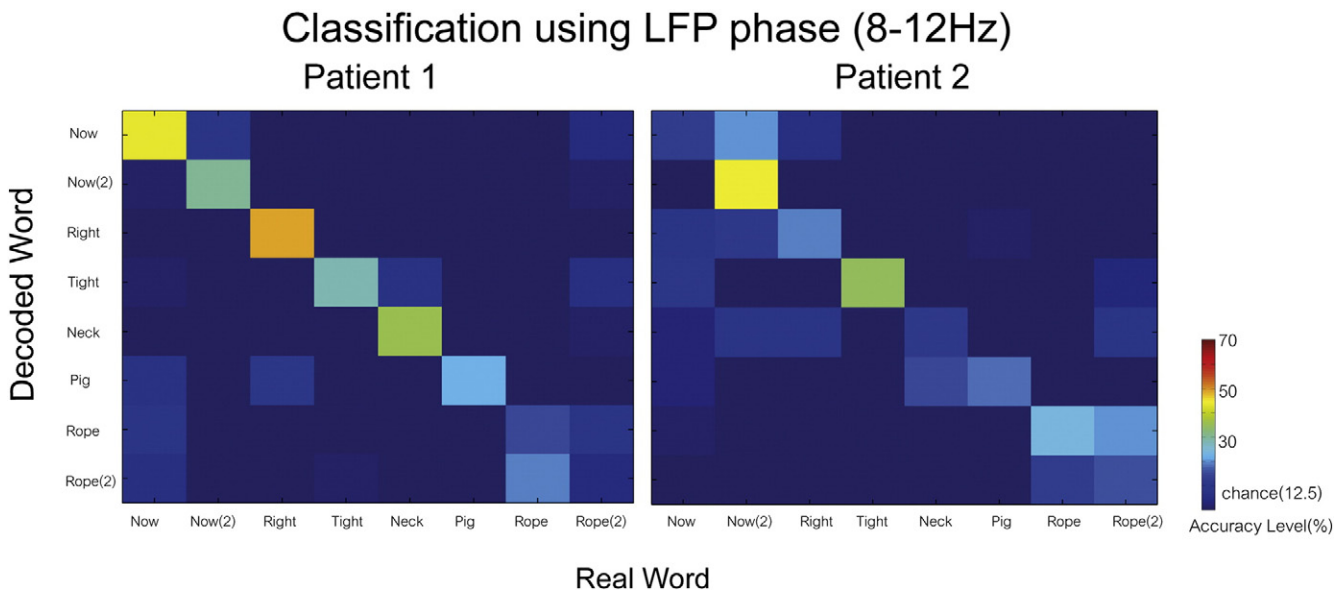

Fig. 6. Word classification performance using LFP. Classification confusion matrix using phase of the LFP signal in the low frequency band (8–12 Hz). Format as in Fig. 4. Color code represents the average proportion of times a trial from a specific word (rows) was assigned to the other words (columns). The diagonal represents correctly decoded words. Theoretical chance level (12.5%) is color coded in dark blue.

used to successfully decode words and serve as a candidate for future brain–machine interface in patients with deficits in speech production.

Acknowledgments

This study was supported by the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation (grant no. 51/11) (R.M. and I.F.); The Israel Science Foundation (grant nos. 1771/13 and 2043/13), and Human Frontiers Science Project (HFSP) Career Development Award (CDA00078/2011-C) (R.M.); The Yosef Sagol Scholarship for Neuroscience Research, The Israeli Presidential Honorary Scholarship for Neuroscience Research, and the Sagol School of Neuroscience Fellowship (O.O.). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank the patients for participating in the study. We also thank E. Ho, T. Fields, and E. Behnke for technical assistance; B. Salaz and I. Wainwright for administrative help.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M., 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. 98 (23), 13367–13372.
Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528.
Bitterman, Y., Mukamel, R., Malach, R., Fried, I., Nelken, I., 2008. Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature 451, 197–201.
Boemio, A., Fromm, S., Braun, A., Poeppel, D., 2005. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat. Neurosci. 8, 389–395.
Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495 (7441), 327–332.
Brumberg, J.S., Nieto-Castanon, A., Kennedy, P.R., Guenther, F.H., 2010. Brain–computer interfaces for speech communication. Speech Comm. 52 (4), 367–379.
Buchsbaum, B.R., Hickok, G., Humphries, C., 2001. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cogn. Sci. 25 (5), 663–678.
Canolty, R.T., Soltani, M., Dalal, S.S., Edwards, E., Dronkers, N.F., Nagarajan, S.S., …, Knight, R.T., 2007. Spatiotemporal dynamics of word processing in the human brain. Front. Neurosci. 1 (1), 185.
Chan, A.M., Dykstra, A.R., Jayaram, V., Leonard, M.K., Travis, K.E., Gygi, B., et al., 2014. Speech-specific tuning of neurons in human superior temporal gyrus. Cereb. Cortex 24 (10), 2679–2693. http://dx.doi.org/10.1093/cercor/bht127.
Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM TIST 2 (3), 27.
Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., Knight, R.T., 2010. Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. 13 (11), 1428–1432.
Chechik, G., Anderson, M.J., Bar-Yosef, O., Young, E.D., Tishby, N., et al., 2006. Reduction of information redundancy in the ascending auditory pathway. Neuron 51, 359–368.
Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., Pesaran, B., 2014. Sensory-motor transformations for speech occur bilaterally. Nature 507 (7490), 94–98. http://dx.doi.org/10.1038/nature12935 [Epub 2014 Jan 15].
Creutzfeldt, O., Ojemann, G., Lettich, E., 1989. Neuronal activity in the human lateral temporal lobe. Exp. Brain Res. 77 (3), 451–475.
Flinker, A., Chang, E.F., Barbaro, N.M., Berger, M.S., Knight, R.T., 2011. Sub-centimeter language organization in the human temporal lobe. Brain Lang. 117 (3), 103–109.
Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322, 970–973.
Fried, I., Wilson, C.L., Maidment, N.T., Engel, J., Behnke, E., et al., 1999. Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. Technical note. J. Neurosurg. 91, 697–705.
Fried, I., Mukamel, R., Kreiman, G., 2011. Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. Neuron 69 (3), 548–562.
Gehr, D.D., Komiya, H., Eggermont, J.J., 2000. Neuronal responses in cat primary auditory cortex to natural and altered species-specific calls. Hear. Res. 150, 27–42.
Grace, J.A., Amin, N., Singh, N.C., Theunissen, F.E., 2003. Selectivity for conspecific song in the zebra finch auditory forebrain. J. Neurophysiol. 89, 472–487.
Guenther, F.H., Brumberg, J.S., Wright, E.J., Nieto-Castanon, A., Tourville, J.A., Panko, M., …, Kennedy, P.R., 2009. A wireless brain–machine interface for real-time speech synthesis. PLoS One 4 (12), e8218.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8 (5), 393–402.
Joris, P.X., Schreiner, C.E., Rees, A., 2004. Neural processing of amplitude-modulated sounds. Physiol. Rev. 84 (2), 541–577.
Kayser, C., Montemurro, M.A., Logothetis, N.K., Panzeri, S., 2009. Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. Neuron 61, 597–608.
Lakatos, P., Chen, C.M., O'Connell, M.N., Mills, A., Schroeder, C.E., 2007. Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53, 279–292.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.
Machens, C.K., Schutze, H., Franz, A., Kolesnikova, O., Stemmler, M.B., et al., 2003. Single auditory neurons rapidly discriminate conspecific communication signals. Nat. Neurosci. 6, 341–342.
Martin, S., Brunner, P., Holdgraf, C., Heinze, H.J., Crone, N.E., Rieger, J., …, Pasley, B.N., 2014. Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroengineering 7.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485 (7397), 233–236.
Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343 (6174), 1006–1010.
Mukamel, R., Fried, I., 2012. Human intracranial recordings and cognitive neuroscience. Annu. Rev. Psychol. 63, 511–537.
Mukamel, R., Nir, Y., Harel, M., Arieli, A., Malach, R., Fried, I., 2011. Invariance of firing rate and field potential dynamics to stimulus modulation rate in human auditory cortex. Hum. Brain Mapp. 32 (8), 1181–1193.

Narayan, R., Grana, G., Sen, K., 2006. Distinct time scales in cortical discrimination of natural sounds in songbirds. J. Neurophysiol. 96, 252–258.
Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., et al., 2007. Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. Curr. Biol. 17, 1275–1285.
Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., …, Brugge, J.F., 2009. Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. 29 (49), 15564–15574.
Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., et al., 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10 (1), e1001251. http://dx.doi.org/10.1371/journal.pbio.100125.
Patel, A.D., Balaban, E., 2000. Temporal patterns of human cortical activity reflect tone sequence structure. Nature 404, 80–84.
Pei, X., Barbour, D.L., Leuthardt, E.C., Schalk, G., 2011. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. J. Neural Eng. 8 (4), 046028.
Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Comm. 41, 245–255.
Quiroga, R.Q., Nadasdy, Z., Ben-Shaul, Y., 2004. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput. 16, 1661–1687.
Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I., 2005. Invariant visual representation by single neurons in the human brain. Nature 435, 1102–1107.
Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., et al., 2001. Probabilistic mapping and volume measurement of human primary auditory cortex. NeuroImage 13 (4), 669–683.
Russ, B.E., Ackelson, A.L., Baker, A.E., Cohen, Y.E., 2008. Coding of auditory-stimulus identity in the auditory non-spatial processing stream. J. Neurophysiol. 99, 87–95.
Schnupp, J.W., Hall, T.M., Kokelaar, R.F., Ahmed, B., 2006. Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. J. Neurosci. 26, 4785–4795.
Steinschneider, M., Volkov, I.O., Fishman, Y.I., Oya, H., Arezzo, J.C., Howard, M.A., 2005. Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter. Cereb. Cortex 15 (2), 170–186.
Tankus, A., Fried, I., Shoham, S., 2012. Structured neuronal encoding and decoding of human speech features. Nat. Commun. 3, 1015.
Wang, X., Kadia, S.C., 2001. Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. J. Neurophysiol. 86, 2616–2620.
Wessinger, C.M., VanMeter, J., Tian, B., Van-Lare, J., Pekar, J., Rauschecker, J.P., 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. J. Cogn. Neurosci. 13 (1), 1–7.
Young, E.D., 2008. Neural representation of spectral and temporal information in speech. Philos. Trans. R. Soc. B-Biol. Sci. 363 (1493), 923–945.
Zheng, Z.Z., Munhall, K.G., Johnsrude, I.S., 2010. Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production. J. Cogn. Neurosci. 22 (8), 1770–1781.
Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., …, Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77 (5), 980–991.